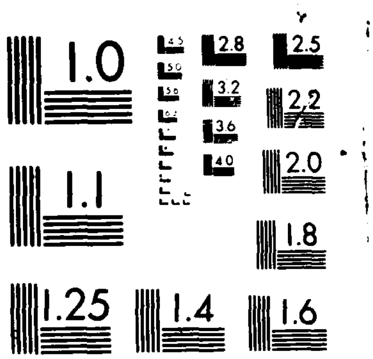
AD-A185 132 FINAL REPORT ON CONTRACT F49620-85-C-0026 VOLUME 4(U) 1/1
PRINCETON UNIV NJ DEPT OF MECHANICAL AND AEROSPACE
ENGINEERING S A ORSZAG MAY 87 AFOSR-TR-87-1349-VOL-4
UNCLASSIFIED F49620-85-C-0026 F/G 20/4 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

AFOSR-TR- 87-1349

(2)

DTIC FILE COPY

AD-A185 132

FINAL REPORT ON AFOSR CONTRACT F49620-85-C-0026

Steven A. Orszag, Principal Investigator
Department of Mechanical and Aerospace Engineering
Princeton University
Princeton, NJ 08544

Volume 4

VET NOV TES TAM EN TVM

DEI SVB NVMINE VIGET

DTIC
S ELECTE D
SEP 3 0 1987
D

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

PRINCETON UNIVERSITY

87 9 24 249

2

FINAL REPORT ON
AFOSR CONTRACT F49620-85-C-0026

Steven A. Orszag, Principal Investigator
Department of Mechanical and Aerospace Engineering
Princeton University
Princeton, NJ 08544

Volume 4

DTIC
ELECTE
SEP 3 0 1987
S D
D

DISTRIBUTION
Approved for
Distribution

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

ADA185-132

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for Public Release; distribution is unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 87-1349
6a. NAME OF PERFORMING ORGANIZATION: Princeton University / 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: AFOSR/NA
6c. ADDRESS (City, State and ZIP Code): Princeton University, Princeton, NJ 08544	7b. ADDRESS (City, State and ZIP Code): Building 410 Bolling AFB DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR/NA / 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: F49620-85-C-0026
8c. ADDRESS (City, State and ZIP Code): Bldg 410, Bolling Air Force Base, Washington, DC 20332-6448	10. SOURCE OF FUNDING NOS. — PROGRAM ELEMENT NO.: 61102F; PROJECT NO.: 2307; TASK NO.: A2; WORK UNIT NO.:
11. TITLE (Include Security Classification): Final Report on Contract F49620-85-C-0026 Vol. 4	

12. PERSONAL AUTHOR(S): Steven A. Orszag

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Final Report	FROM 10/1/84 TO 11/30/86	May, 1987	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	Turbulence, Numerical Simulation

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report consists of papers that summarize work done on this research project. The major results include: 1) The development and application of the renormalization group method to the calculation of fundamental constants of turbulence, the construction of turbulence transport models, and large-eddy simulations; 2) The application of RNG methods to turbulent heat transfer through the entire range of experimentally accessible Reynolds numbers; 3) The discovery that high Reynolds number turbulent flows tend to act as if they had weak nonlinearities, at least when viewed in terms of suitable 'quasi-particles'; 4) The further analysis of secondary instability mechanisms in free shear flows, including the role of these instabilities in chaotic, 3-D free shear flows; 5) The further development of numerical simulations of turbulent spots in wall bounded shear flows; 6) The study of cellular automata for the solution of fluid mechanical problems; 7) The clarification of the relationship between the hyperscale instability of anisotropic small-scale flow structures to long-wavelength perturbations and the cellular automaton description →

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☒ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION: Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL: Dr James M McMichael	22b. TELEPHONE NUMBER (Include Area Code): (202) 767-4936 / 22c. OFFICE SYMBOL: AFOSR/NA

DD FORM 1473, 83 APR — EDITION OF 1 JAN 73 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

STRACT, continued from other side

fluids; 8) The development of efficient methods to analyze the structure of range attractors in the description of dynamical systems; 9) The analysis of perscale instability as a mechanism for destabilization of coherent flow structures.

WEAK INTERACTIONS AND LOCAL ORDER IN STRONG TURBULENCE

Victor Yakhot

Steven A. Orszag

Alexander Yakhot[(a)]

Raj Panda

Applied and Computational Mathematics
Princeton University
Princeton, NJ 08544
U.S.A.

Uriel Frisch

CNRS
Observatoire de Nice
B. P. 139, 06003 Nice Cedex, FRANCE

and

Robert H. Kraichnan[(b)]

303 Potrillo Drive
Los Alamos, NM 87544
U.S.A.

By
Dist
A-1

November 3, 1986
Revised April, 1987

ABSTRACT

Data from simulations of channel flow and decay of homogeneous turbulence indicate anomalously strong correlation of velocity and vorticity directions ('local Beltramization') in band-filtered velocity fields when the band consists of a thin cigar in mode space (whose physical space representation is as an array of 'pancake eddies'). Spherical shells or other broad bands in mode space do not seem to exhibit the effect.

In fully developed three-dimensional turbulence, 'local' interactions (that is, interactions between comparable scales) have been traditionally thought to be dynamically much more significant than nonlocal interactions (for example, in the cascade of turbulent energy).[1]

However, analysis of experimental data on energy transfer shows[2] that those interactions which contribute most strongly involve wavenumber triads which, while local, have aspect ratios (ratio of largest to smallest wavenumber) in the range 5 to 10. The ALHDIA[3] yields results typical of second-order turbulence. It predicts that about 20% of the total transfer come from triads interactions with aspect ratios less than 2, while about 80% come from triads with aspect ratios less than eight. However, these figures give less than the whole story. If local interactions are consistently *removed* from the dynamics, the loss of the energy transfer from these interactions is counterbalanced with an enhanced transfer by the remaining interactions, at least within the framework of the closures[5]. The result is that removing triads with aspect ratios eight and less leaves the total energy transfer nearly unchanged, instead of reducing it by 80%.

To analyze the effect of local versus nonlocal interactions, we decompose the velocity field $\mathbf{v}(\mathbf{r})$ and the vorticity field $\boldsymbol{\omega} = \nabla \times \mathbf{v}$ as $\mathbf{v}(\mathbf{r}) = \sum \mathbf{v}^j(\mathbf{r})$; $\boldsymbol{\omega}(\mathbf{r}) = \sum \boldsymbol{\omega}^j(\mathbf{r})$. Here the band-filtered velocity field $\mathbf{v}^j(\mathbf{r})$ is defined as $\mathbf{v}^j(\mathbf{r}) = \sum \mathbf{u}(\mathbf{k})\exp(i\mathbf{k}\cdot\mathbf{r})$ where the sum extends over wavevectors $\mathbf{k}$ within suitable distinct wavevector bands B_j. More generally, the band filtered field is defined as $\mathbf{v}^j(\mathbf{r}) = \sum_{n \in B_j} \mathbf{u}(n)\phi_n(\mathbf{r})$ where $\phi_n(\mathbf{r})$ is a complete set of orthogonal functions. In the channel flow simulations reported below, the velocity field is Fourier transformed in the x- and y-directions while the Chebyshev polynomials $T_p(z)$ are used in the z-direction perpendicular to the walls. In the simulations of decaying turbulence reported below, three-dimensional Fourier series are used.

In terms of the band-filtered fields, the Navier-Stokes equations for incompressible flow are

$$\frac{\partial \mathbf{v}}{\partial t} = \sum_{i,j} N_{i,j} + \nu_0 \nabla^2 \mathbf{v} \qquad (1)$$

where $N_{i,j} = \mathbf{v}^i \times \omega^j - \nabla \pi_{i,j}$ and $\nabla \pi_{i,j}$ subtracts out the divergent part of $\mathbf{v}^i \times \omega^j$. A measure of the strength of local interactions is given by the intra-band contribution $N_{i,i}$ to the nonlinear term in (1).

In two-dimensional flow, velocity and vorticity are perpendicular, so the intra-band term $N_{i,i}$ can only be small through pressure balancing. This is indeed observed in high-Reynolds number numerical simulations of decaying two-dimensional flow, which display organization of the flow into quasi-one-dimensional vortex-gradient sheets[8] or nearly circular vortex patches[9].

In three-dimensional flow, depletion of local nonlinear interactions can occur either by pressure balance within $N_{i,i}$ or, more directly, by local 'Beltramization' in which the band-filtered fields $\mathbf{v}^i$ and ω^i are nearly parallel in physical space. While local pressure balance does not allow for complicated instantaneous three-dimensional flow topologies,[10] local Beltramized flows are not so restricted. In this work, we have investigated the possibility of local Beltramization using results of a direct numerical of turbulent channel flow at a Reynolds number $Re_* = 184$ based on the half-channel-width and the friction velocity. Thus the channel width in dimensionless wall coordinates is $H = 368$. The computer code is described in Ref. 11. Two channels of different linear dimension have been considered. The number of grid points used in these simulations was 32^3 and $32^2 \times 64$, respectively. The same effect has also been studied in (128^3) simulations of the decay of homogeneous turbulence.

The analysis used here follows an earlier analysis of the full (non-bandpass-filtered) flow field $v(\mathbf{r})$.[12] Specifically, we evaluate the distribution of the cosine of the angle θ between the velocity $v^i(\mathbf{r})$ and vorticity $\omega^i(\mathbf{r})$. We subdivide the interval $-1 \leq \cos\theta \leq 1$ into two hundred equally spaced intervals and define the probability function $P(\cos\theta)$ as the number of grid points having the angle θ between the bandpass-filtered velocity and vorticity. The variable $\cos\theta$ is the relevant one in three-dimensions because of the form of the volume element. Note that $\sum P(\cos\theta) = N$ where N is the total number of grid points. Each half of the channel ($0 < z < 184$) is subdivided into three intervals: $0 \leq z \leq 15$; $15 \leq z \leq 40$; $40 \leq z \leq 184$. In order to avoid the effects of the walls, the velocity field has been analyzed in the outer part of the channel $40 \leq z \leq 328$. The Fourier-Chebyshev band-filtering has been performed using the data on the entire flow field. Then the filtered field is transformed back into physical space and the resulting field is analysed locally in space for $40 \leq z \leq 328$. We recognize that there may be ambiguity in space localization due to an "uncertainty principle".

The results plotted in Fig. 1 for the band B_i consisting of $k_x = \pm 4$; $k_y = \pm 2$; $p = 15, 16, 17, 18$ show that $v^i(r)$ and $\omega^i(r)$, tend to align; that is the probability function $P(\cos\theta)$ is sharply peaked at $\cos\theta = -1$ at $t = 50$. To characterize the effect we define the Beltrami ratio as $B = \max[P(1), P(-1)]/P_{av}(0)$, where $P_{av}(0)$ is the value of $P(\cos\theta)$ averaged over the interval $-\frac{1}{3} < \cos\theta < \frac{1}{3}$. In Fig. 2, we plot the time evolution of $B(t)$ for $40 < t < 77$ with samples taken at the time intervals $\Delta t = 1$. We find that $B > 12$ approximately 30% of the time. Continuing to add modes to the band leads to disappearance of the peak of $P(\cos\theta)$ when the band is too wide. A similar effect has been observed in numerical simulation of decaying homogeneous turbulence (see Fig. 3). It is typical for a variety of realizations of band-filtered velocity fields in both channel and decaying turbulence simulations that as soon as a few modes are present, there is an organization of the flow field in which velocity and vorticity are almost aligned.

It should be mentioned that the band-filtered velocity field v^j does not satisfy the incompressibility condition $\nabla \cdot v^j = 0$ because of the properties of the Chebyshev polynomials. To assess the role of incompressibility we have subtracted the nonsolenoidal part from the filtered velocity field and calculated the probability function $P(\cos\theta)$ of the residual field satisfying the $\nabla \cdot v = 0$ constraint. The results are presented in Fig. 1(b), 2(b). It is obvious that imposition of the incompressibility condition does not strongly change the results presented in Fig. 1(a), 2(a) since the filtered velocity field $v(k_x,k_y,p)$ is only weakly compressible in the part of the channel we consider here ($z_+ > 40$). It is only in the wall region $z_+ < 15$ that the effect is not observed. It is clear from Fig. 2(b) that imposing the incompressibility constraint leads to an increase of the value of B by some 10-40%. An interesting property of the probability density $P(\cos\theta)$ is its asymmetry $P(1) \neq P(-1)$ when the Beltrami ratio B is large. However, the mean distribution $\overline{P(\cos\theta)}$ obtained from 35 realizations in the interval $40 \leq t < 75$ is very symmetric.

This 'local Beltramization' effect is observed to occur for general wavevector packets W with large aspect ratio $k >> \Delta k >> 1$ where k is a typical wavenumber in the band and Δk is the 'diameter' of the packet [$\Delta k = \max(|\mathbf{k} - \mathbf{p}|$ with $\mathbf{k}, \mathbf{p} \in W)$]. These packets produce 'pancake' eddies in which both the velocity and vorticity are quasi-two dimensional fields. This may be most easily seen by considering a prototype of such a packet, namely, one for which $k_z << k_x,k_y$ for all $\mathbf{k}$ in the packet and k_z restricted to a narrow range. Thus the incompressibility condition gives $w = -(k_x u + k_y v)/k_z << |u| + |v|$, while the vorticity field $\boldsymbol{\omega} = i\mathbf{k}\times\mathbf{v}$ satisfies $\omega_z << |\omega_x| + |\omega_y|$. For such pancake eddies, we would expect, in the absence of other correlations, that the angle between v and $\boldsymbol{\omega}$ is uniformly distributed in the plane of the pancake, so that $P(\cos\theta) \propto 1/|\sin\theta|$ with strong peaks at $\cos\theta = \pm 1$.

To test whether the observed effect (see Fig. 1, 2) is of dynamic or kinematic nature we chose a number of distinct band-filtered fields corresponding to a few different instants of time t (see Fig. 2(a)). From these fields we generated 1000 different realizations by uniformly distributing the complex numbers $v(k_x,k_y,p)$ for each selected mode in the band. Imposing the incompressibility condition on each of the band-filtered random fields we have generated another 1000 realizations. An additional 500 random realizations were generated by uniformly distributing the random complex numbers to each of the $(32\times 32\times 64)$ modes of the total field $v(k_x,k_y,p)$. The resulting random velocity fields were made incompressible and then band filtered. The weakly compressible band-filtered fields were again made incompressible by subtraction of the nonsolenoidal part. Then the probability density $P(\cos\theta)$ was calculated for each member of the ensemble consisting of 2500 realizations. We then calculated the cumulative probability p(B) that gives the probability that one Beltrami ratio is larger than B. The function p(B) derived from these 2500 fields was compared with the function p(B) corresponding to 35 realizations obtained from the time evolution of the band-filtered numerical solution of the Navier-Stokes equation at $40 \leq t \leq 75$ taken at $\Delta t = 1$. The results are presented in Fig. 4. These runs suggest that the observed effect does not have a simple kinematic explanation. However we must note that the Chebyshev-Fourier modes do not diagonalize the two-point covariance function of the observed velocity field and for this reason the randomized comparison fields do not exhibit the two-point covariance of the observed fields. This property will be investigated later[13].

The generality of the above results on the bandpass-filtered geometrical order of velocity and vorticity is yet to be tested on other turbulent flows. It is difficult now to give a systematic explanation for local Beltramization. This may actually be as, or more, difficult than explaining the remarkable buildup of correlations between velocity and magnetic fields in strong MHD turbulence.[14] We stress that it is unlikely that any of the traditional two-point closures can provide insight into the process of local Beltramization.[15]

We now speculate on several instability mechanisms that may be important in the dynamics that lead to local Beltramization.[19] First, secondary instability is a short wavelength instability of quasi-two-dimensional flows that tends to produce locally aligned vorticity and velocity. On the other hand, anisotropic pancake eddies are subject to the long-wavelength hyperscale instability. These eddies are also unstable to the AKA instability[18] which also leads to generation of Beltrami flows at large scales. Perhaps a combination of these instabilities, acting cyclically, can explain the observed effects.

The importance of steady solutions of the Euler equation in the context of turbulence has been discussed by H.K. Moffat[20]. he argues that turbulent flows can be represented in terms of coherent structures with large helicity (Beltrami flows).[20] Our findings support Moffatt's conjecture in the sense that the flow may be a superposition of Beltrami-like structures defined at many different scales.

Whatever the cause of local Beltramization, it can have far reaching consequences. Indeed, we may conjecture that fully developed turbulence rearranges itself into a hierarchy of coherent near-Beltrami flows with minimal self-interaction. In this case, turbulence should be amenable to multiple-scale perturbation theory[21] or renormalization-group methods. Furthermore, virtually all turbulence models, starting with those of Prandtl, are based on scale separation to justify eddy viscosity concepts. The present ideas seem to provide some justification for these approaches.

References

(a) On leave from Department of Mechanical Engineering, Ben-Gurion University of the Negev, Beer-Sheva, ISRAEL.

(b) Consultant, Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory.

1. A.S. Monin and A.N. Yaglom, Statistical Fluid Mechanics, Vol. 2, M.I.T. Press, 1975.
2. R.G. Deissler, Appl. Sci. Res. **34**, 379 (1978).
3. R.H. Kraichnan, Phys. Fluids **9**, 1728 (1966).
4. R.H. Kraichnan, Phys. Fluids (submitted).
5. M.E. Brachet, M. Meneguzzi, and P.-L. Sulem, Phys. Rev. Lett. **57**, 683 (1986).
6. B. Fornberg, J. Comp. Phys. **25**, 1 (1977); C. Basdevant, B. Legras, R. Sadourny and M. Belian, J. Atmos. Sci. **38**, 2305 (1981); C. McWilliams, J. Fluid Mech. **146**, 21 (1984).
7. V. Arnold, Comptes Rendus Acad. Sci. (Paris) **261**, 17 (1965).
8. S.A. Orszag and A. Patera, Phys. Rev. Lett. **47**, 832 (1981).
9. R. Pelz, V. Yakhot, S.A. Orszag, E. Levich and L. Shtilman, Phys. Rev. Lett. **54**, 2505 (1985).
10. There is insufficient data on the dynamics of the band-filtered fields in decaying turbulence. Thus, the data plotted in Fig. 3 can serve only as a preliminary indication of the possibility of the local Beltramization effect in decaying homogeneous turbulence.
11. A. Pouquet, M. Meneguzzi and U. Frisch, Phys. Rev. **A33**, 4266 (1986).
12. A simple model which shows the limitations of two-part closures in the MHD context is given by U. Frisch, A. Pouquet, P.-L. Sulem and M. Meneguzzi, J. Mec. Th. Appl. (Paris), special 1983 issue "Two-Dimensional Turbulence", p. 191.
13. V. Yakhot and G. Sivashinsky, Phys. Rev. A **35**, 815 (1987).
14. V. Yakhot and R. Pelz, Phys. Fluids (in press).
15. U. Frisch, Z. She and P.-L. Sulem, Physica D, submitted.
16. B.J. Bayly and S.A. Orszag, Ann. Rev. Fluid Mech., 1988 (in press).
17. H.K. Moffatt, J. Fluid Mech. **166**, 359 (1986).
18. A. Yoshizawa, Phys. Fluids **25**, 1532 (1982); **27**, 1377 (1984).

19. V. Yakhot and S.A. Orszag, Phys. Rev. Lett. **57**, 1722 (1986); J. Sci. Comp. **1**, 3 (1986).

FIGURE CAPTIONS

Fig. 1 Probability distribution $P(\cos\theta)$ in the channel flow in the region $40 < z_+ < 328$ for:

a. the band-filtered velocity field ($k_x = \pm 2; k_y = \pm 2; P = 15,16,17,18$)

b. the same band-filtered field with the incompressibility condition $\nabla \cdot v^j = 0$ imposed.

Fig. 2 Time evolution of the Beltrami ratio for the band-filtered field:

Solid line: velocity field $v^j(r)$, ($k_x = \pm 2; k_y = \pm 2; P = 15,16,17,18$)

Dotted line: the same field with the incompressibility condition imposed.

Fig. 3 Probability function $P(\cos\theta)$ for the band-filtered field v^j ($k_x = \pm 2; k_y = \pm 2; k_2 = \pm 22$) in a simulation of decaying, homogeneous turbulence.

Fig. 4 Cumulative probabilities p(B):

Curve I (•): corresponds to 31 realizations of $v(r)$ ($k_k = \pm 2; k_y = \pm 2; p = 15-18$) taken from the time evolution of the flow.

Curve II (•): p(B) generated by introducing random phases into the fields $v^j(r)$.

Curves (×): same as (•) but with the incompressibility imposed on all fields $v^j(r)$.

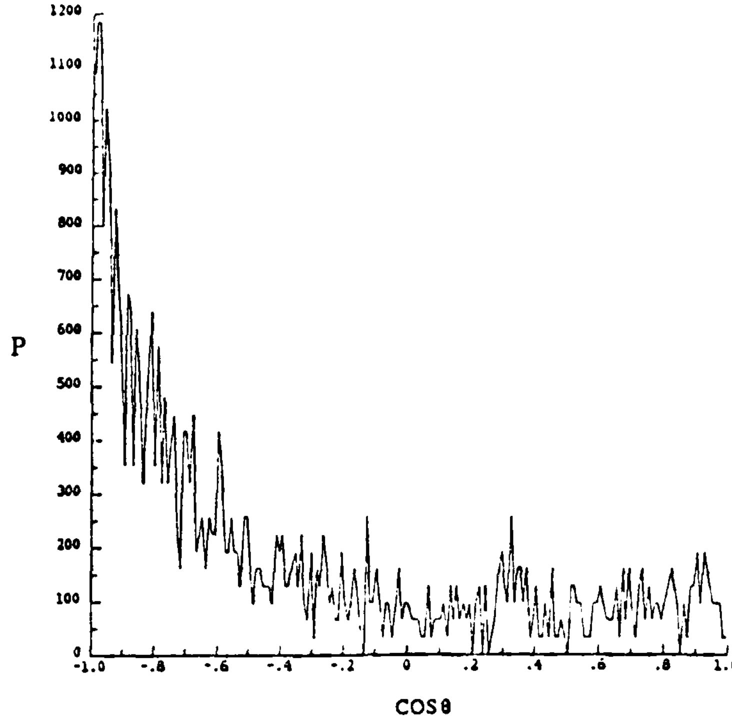


Figure 1a

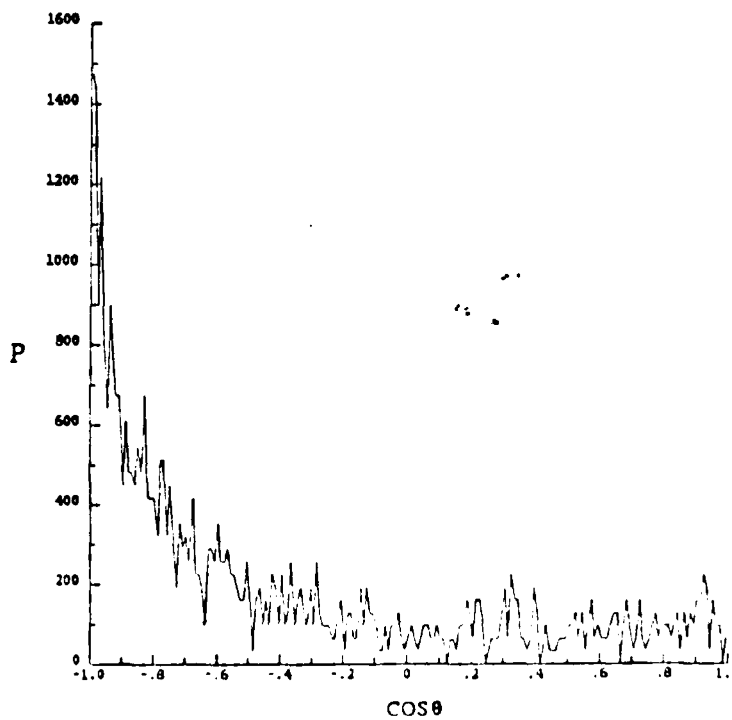


Figure 1b

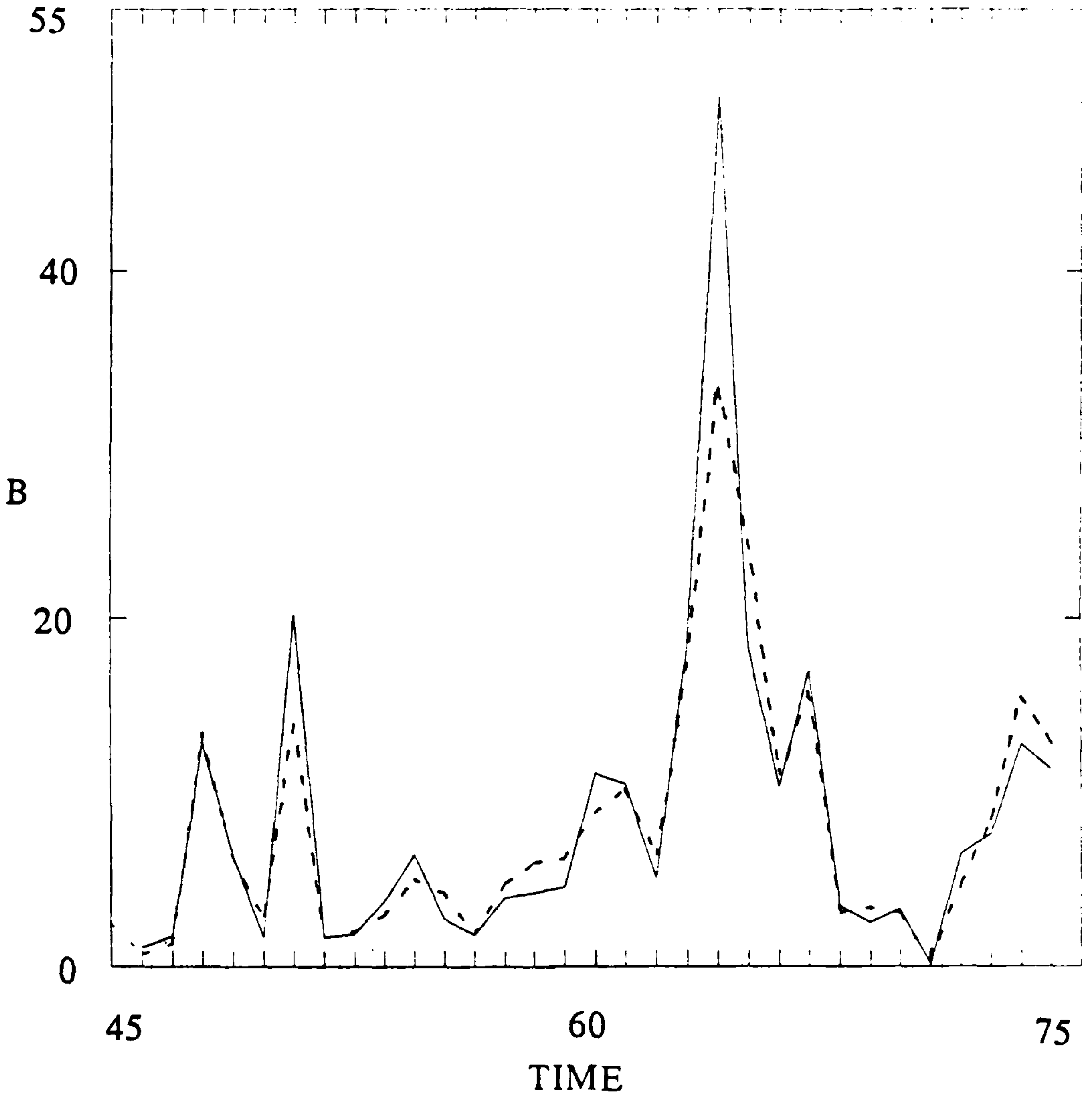
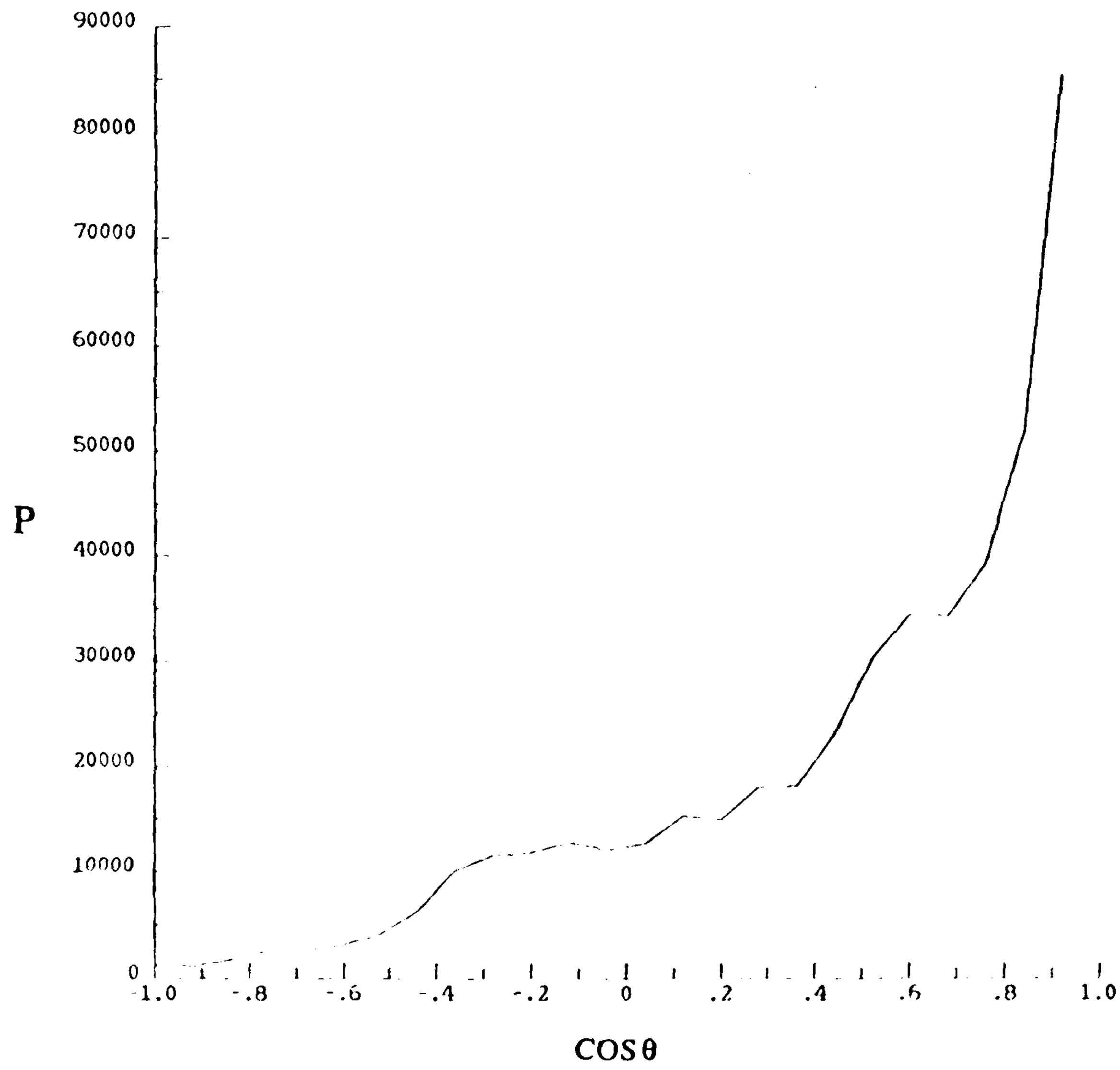


Figure 2



P
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
-1.0
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1.0
COSθ

Figure 3

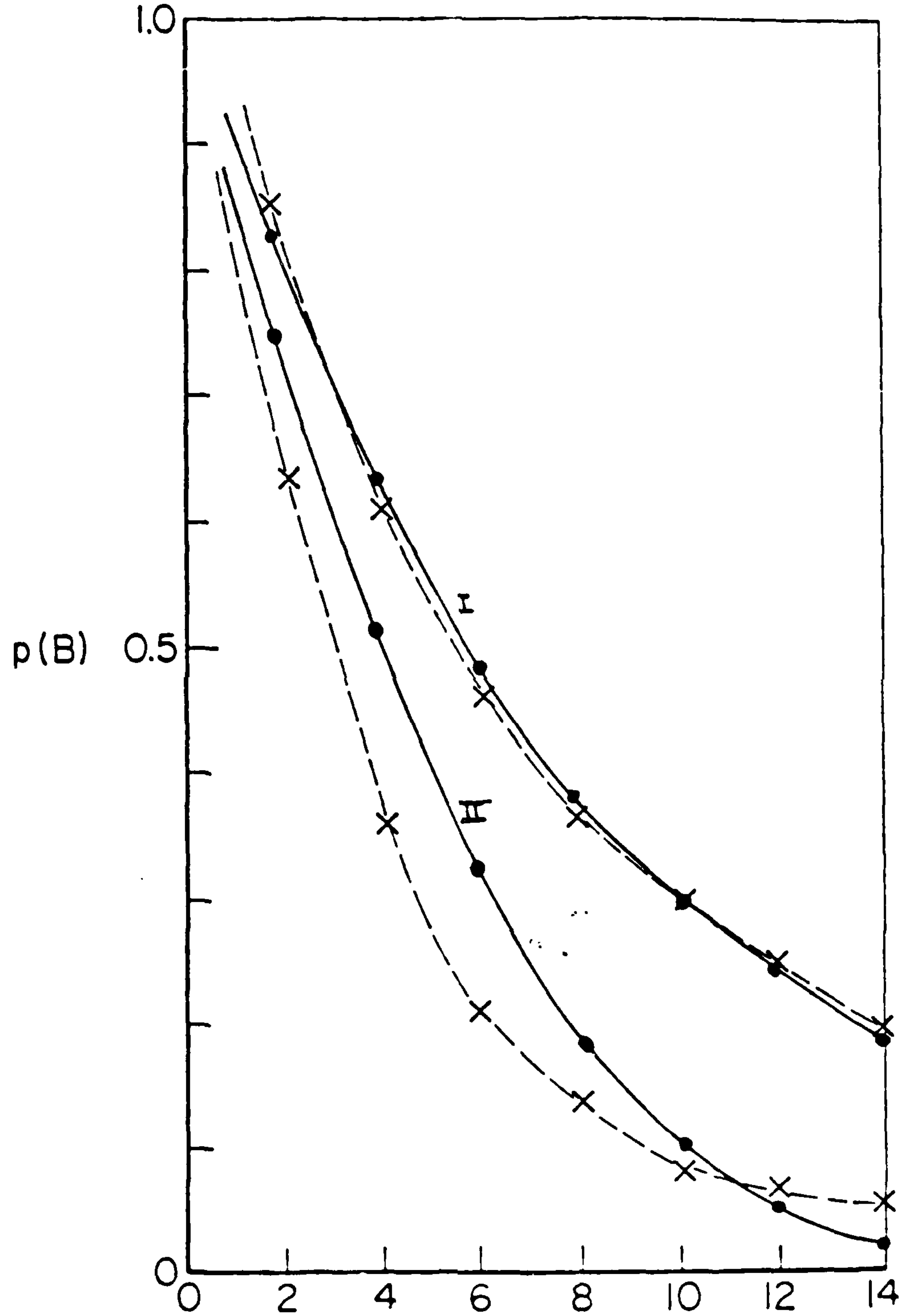


Figure 4

The purpose of this Letters section is to provide rapid dissemination of important new results in the fields regularly covered by The Physics of Fluids. Results of extended research should not be presented as a series of letters in place of comprehensive articles. Letters cannot exceed three printed pages in length, including space allowed for title, figures, tables, references and an abstract limited to about 100 words.

Relation between the Kolmogorov and Batchelor constants

Victor Yakhot and Steven A. Orszag
Applied and Computational Mathematics, Princeton University, Princeton, New Jersey 08544

(Received 2 September 1986; accepted 23 October 1986)

It is shown that the renormalization group theory of turbulence leads to the relation $\mathrm{Ba} = C_K P_t$ between the turbulent Prandtl number P_t, the Kolmogorov constant C_K, and the Batchelor constant Ba.

The purpose of this Letter is to show that the renormalization group theory[1,2] of turbulence leads to a simple relation between the Batchelor (Ba) and Kolmogorov (C_K) constants. Here Ba and C_K are defined by the inertial range kinetic energy and passive scalar spectra [$E(k)$ and $E_T(k)$, respectively]:

$$E = C_K \bar{\epsilon}^{2/3} k^{-5/3} \tag{1}$$

and

$$E_T = \mathrm{Ba}(N/\bar{\epsilon}^{1/3}) k^{-5/3}. \tag{2}$$

The parameters N and $\bar{\epsilon}$ are the scalar and kinetic energy dissipation rates defined as

$$N = \kappa_0 \overline{\left(\frac{\partial T}{\partial x_i}\right)^2} \tag{3}$$

and

$$\bar{\epsilon} = \frac{\nu_0}{2} \overline{\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)^2} \tag{4}$$

and κ_0 and ν_0 are molecular diffusivity and viscosity, respectively.

In the first step of the renormalization group procedure, we eliminate modes with wavenumbers larger than p from the equation of motion for the Fourier components of the velocity and scalar fields $v_1(\mathbf{k},\omega)$ and $T(\mathbf{k},\omega)$, respectively. In this calculation, it is assumed that $k \ll p$. In the limit $k \ll p$, the sole effect of the small-scale elimination procedure is the generation of a tubulent viscosity $\nu(p)$ and diffusivity $\kappa(p)$. In the limit $k \ll p \ll k_d$, where k_d is the Kolmogorov dissipation wavenumber, the turbulent transport coefficients are proportional and

$$\nu(p) = P_t \kappa(p) \tag{5}$$

with $P_t = 0.7179$.

We define the velocity and scalar fields v_i^p and T^p following the elimination of wavenumbers larger than p. It can be shown[1,2] that the two scalars

$$\bar{\epsilon}(p) = \frac{\nu(p)}{2} \overline{\left(\frac{\partial v_i^p}{\partial x_j} + \frac{\partial v_j^p}{\partial x_i}\right)^2} \tag{6}$$

and

$$N(p) = \kappa(p) \overline{\left(\frac{\partial T_i^p}{\partial x_j}\right)^2} \tag{7}$$

are constants independent of p. This has some important implications. If the spectra are given by relations (1) and (2) then

$$\bar{\epsilon} = 2\nu(p) \int_0^p k^2 E(k) dk = \frac{3}{2} \nu(p) C_K \bar{\epsilon}^{2/3} p^{4/3} \tag{8}$$

and

$$N = 2\kappa(p) \int_0^p k^2 E_T(k) dk = \frac{3}{2} \kappa(p) \frac{N}{\bar{\epsilon}^{1/3}} \mathrm{Ba}\, p^{4/3}. \tag{9}$$

Dividing (8) by (9) and using relation (5) we obtain

$$\mathrm{Ba} = C_K P_t. \tag{10}$$

Substituting the values $C_K = 1.617$ and $P_t = 0.7179$ derived in Refs. 1 and 2 we obtain $\mathrm{Ba} = 1.161$ in good agreement with experimental data.

This result is obtained using the ϵ-expansion in the vicinity of the fixed point. Thus, relation (9), is, strictly speaking, valid in the limit of Reynolds number $R \to \infty$. We believe that formula (9) is accurate when molecular diffusion is negligible in comparison with turbulent diffusion. It should be mentioned that experimental data on Ba are much more scattered than data on the Kolmogorov constant C_K. This can be easily explained since, in many cases, the contribution from even weak natural convection can strongly influence the heat transfer but not a momentum transfer in turbulent flow when R is not sufficiently large.

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research under Contract No. N00014-82-C-0451 and the Air Force Office of Scientific Research under Contract No. F-49620-85-C-0026.

[1] V. Yakhot and S. A. Orszag, Phys. Rev. Lett. 57, 1722 (1986).
[2] V. Yakhot and S. A. Orszag, J. Sci. Comput. 1, 3 (1986).

 0031-9171/87/010003-01$01.90 © 1987 American Institute of Physics

AN EFFICIENT METHOD FOR COMPUTING LEADING EIGENVALUES AND EIGENVECTORS OF LARGE ASYMMETRIC MATRICES

I. Goldhirsch*, Steven A. Orszag & B. K. Maulik
Applied & Computational Mathematics
Princeton University
Princeton, NJ 08544
U.S.A.

*Permanent Address:

Dept. of Fluid Mechanics and Heat Transfer
Faculty of Engineering
Tel-Aviv University
Ramat Aviv, Tel-Aviv 69978
ISRAEL

ABSTRACT

An efficient method for computing a given number of leading eigenvalues (i.e. having largest real parts) and the corresponding eigenvectors of a large asymmetric matrix $\mathbf{M}$ is presented. The method consists of three main steps. The first is a filtering process in which the equation $\dot{x} = \mathbf{M}x$ is solved for an arbitrary initial condition $x(0)$ yielding: $x(t) = e^{\mathbf{M}t}x(0)$. The second step is the construction of $(n + 1)$ linearly independent vectors $v_m = \mathbf{M}^m x$, $0 \leq m \leq n$ or $v_m = e^{m\mathbf{M}\tau}x$ (τ being a "short" time interval). By construction, the vectors v_m are combinations of only a small number of leading eigenvectors of $\mathbf{M}$. The third step consists of an analysis of the vectors $\{v_m\}$ that yields the eigenvalues and eigenvectors.

The proposed method has been successfully tested on several systems. Here we present results pertaining to the Orr-Sommerfeld equation. The method should be useful for many computations in which present methods are too slow or necessitate excessive memory. In particular, we believe it is well suited for hydrodynamic and mechanical stability investigations.

I. INTRODUCTION

Numerous problems in science and technology involve the computation of leading eigenvalues (and sometimes of the corresponding eigenvectors) of large asymmetric matrices. In some cases, such as that of the calculation of eigenvalues of transfer matrices in lattice problems (Stanley, 1971) or in hydrodynamic stability analyses (Drazin & Reid, 1981), the matrix of interest is infinite, in principle.

While rather efficient methods exist for the diagonalization of symmetric matrices (see, e.g. Parlett, 1980 or Lewis, 1977) — no satisfactory algorithm for the asymmetric case is known to us. The main obstacle is, of course, the nonorthogonality of the eigenvectors. The Arnoldi (1951) method and its variants enable the calculation of some eigenvalues, but not necessarily those with the largest real parts (which are important in stability investigations; see however Jenning & Stewart, 1975 and Saad, 1980 and references therein). At best this method produces eigenvalues of *largest moduli.* Another disadvantage of the Arnoldi method is the necessity to increase the dimension of the Krylov (Wilkinson, 1965) subspace considered in order to improve accuracy. This may create memory problems.

Another method (Bennetin *et al*, 1980; Shimada & Nagashima, 1979), which was designed to compute Liapunov exponents for dynamical systems, is capable of producing the real parts of the leading eigenvalues. Its advantage is in the fact that its implementation does not require a large memory. However the method is slowly converging and produces neither the imaginary parts of the eigenvalues nor the eigenvectors. (However, see Goldhirsch *et al*, 1987, for an accelerated convergence method and computation of eigenvectors. The latter can be computed efficiently for relatively small systems.)

It thus seems that many important problems involving large asymmetric matrices are very difficult to analyze using available methods. The algorithm proposed in this paper is designed for such problems. It is *fast* and *simple* and can therefore be easily implemented.

The proposed method has three essential steps. The first step is designed to filter out nonleading eigenvectors. Let **x** be an "initial" vector and M the matrix whose leading eigenvalues are sought. We first solve the equation:

$$\frac{d\mathbf{x}}{dt} = \mathbf{M}\mathbf{x} \tag{1.1}$$

for $0 \leq t \leq t_0$. The resulting vector $\mathbf{x}(t_0) = e^{Mt_0}\mathbf{x}$ is obviously a combination of eigenvectors corresponding to leading eigenvalues (henceforth called leading eigenvectors). The nonleading eigenvectors are being damped by exponential factors. Moreover, if it so happens that the chosen initial vector **x** is independent of a leading eigenvector, then this eigenvector will be introduced into x(t) by roundoff errors in the process of numerical integration of eqn. (1.1). The next step involves the creation of n linearly independent vectors. This can be done either by computing $\{M^m x(t_0);\ 0 \leq m \leq n-1\}$ or by computing $\{e^{mM\tau}x(t_0);\ 0 \leq m \leq n-1\}$. It is obvious (and shown below) that the n vectors created this way are independent for nondegenerate spectra. A method for degenerate spectra will be described below as well. Once we obtain n independent vectors which are essentially linear combinations of leading eigenvectors only, it is a matter of straightforward algebra to produce the leading eigenvalues and eigenvectors.

The structure of the paper is as follows. Section II presents a description of the method we propose. Section III offers error estimates and analyses of the algebraic structure of the vectors produced in the second step of the method. Section IV presents two algorithms based on the proposed method. Section V presents results obtained from an implementation of the method in the case of the Orr-Sommerfeld equation. Section VI offers a brief summary. Those readers not interested in the details of the mathematical analysis of the method may skip Sec. III.

II. DESCRIPTION OF THE METHOD

Consider a matrix **M**, which is, in general, asymmetric and large (say of rank R). Assume that the eigenvalues of **M**: $\lambda_1, \lambda_2, \lambda_3 \cdots$ are arranged so that $\mathrm{Re}\lambda_1 > \mathrm{Re}\lambda_2 > \mathrm{Re}\lambda_3 \cdots$, i.e. it is assumed that the real part of the spectrum of **M** is nondegenerate. Let $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \cdots$ be the right eigenvectors of **M** corresponding to the eigenvalues $\lambda_1, \lambda_2, \lambda_3 \cdots$, respectively.

Any vector **x** can be expanded in terms of the right eigenvectors of **M**:

$$\mathbf{x} = \sum_{i=1}^{R} \alpha_i \mathbf{e}_i \tag{2.1}$$

Upon applying the operator $e^{\mathbf{M}t}$, $t > 0$, to **x** we obtain:

$$e^{\mathbf{M}t}\mathbf{x} = \sum_{i=1}^{R} \alpha_i e^{\lambda_i t} \mathbf{e}_i \tag{2.2}$$

Defining the resulting vector in (2.2) as $\mathbf{x}(t)$, we observe that the vector:

$$\hat{\mathbf{x}}(t) = \frac{\mathbf{x}(t)}{|\mathbf{x}(t)|} \tag{2.3}$$

can be approximated, for large enough t, by:

$$\hat{\mathbf{x}}(t) \simeq \frac{1}{|\mathbf{x}(t)|} \sum_{i=1}^{n} \alpha_i e^{\lambda_i t} \mathbf{e}_i \tag{2.4a}$$

or

$$\hat{\mathbf{x}}(t) \simeq \sum_{i=1}^{n} \beta_i \mathbf{e}_i \tag{2.4b}$$

where

$$\beta_i = \frac{\alpha_i e^{\lambda_i t}}{|x(t)|} \tag{2.4c}$$

and n is a suitably chosen (truncation) integer. The error involved in the approximation (2.4) is of the order $\left|e^{(\lambda_{n+1}-\lambda_1)t}\right|$. Thus, for a given desired accuracy there are values of n and of t which fulfill these requirements. The way t and n are to be chosen in a practical algorithm is explained below. In the rest of this section it is assumed that (2.4) is an actual equality in order to simplify the presentation.

Consider the following vectors:

$$\mathbf{v}_m = \mathbf{M}^{m-1}\hat{\mathbf{x}}(t) \tag{2.5}$$

where $1 \leq m \leq n+1$. Using (2.4b), we obtain:

$$\mathbf{v}_m = \sum_{i=1}^{n} \lambda_i^{m-1}\beta_i e_i; \quad 1 \leq m \leq n+1 \tag{2.6}$$

Consider now the matrix W_{ij} defined as:

$$W_{ij} = \lambda_j^{i-1} \qquad 1 \leq i, j \leq n \tag{2.7}$$

and denote $\beta_i e_i = \bar{e}_i$. The vectors $\bar{e}_i$, $1 \leq i \leq n$ are n linearly independent vectors since the e_i's obviously are. Eqn. (2.6) can be rewritten as:

$$\mathbf{v}_m = \sum_{k=1}^{n} W_{m,k}\bar{e}_k \tag{2.8}$$

The matrix $\mathbf{W}$ is a Vandermonde matrix whose determinant is $\prod_{i \neq j}(\lambda_i - \lambda_j)$ and is, by assumption, nonvanishing. Consequently the n vectors $\mathbf{v}_1, \cdots \mathbf{v}_n$ are independent and spanned by $\mathbf{e}_1, \mathbf{e}_2, \cdots \mathbf{e}_n$. Hence $\mathbf{v}_{n+1}$, which is also spanned by $\mathbf{e}_1, \cdots, \mathbf{e}_n$, can be written as:

$$v_{n+1} = \gamma_1 v_1 + \gamma_2 v_2 + \cdots \gamma_n v_n \tag{2.9}$$

Consider now the following linear transformation defined by:

$$Tv_m = v_{m+1} \quad 1 \leq m \leq n \tag{2.10}$$

This is a linear transformation from the subspace spanned by $v_1, \cdots v_n$ or $e_1, \cdots e_n$ to itself. Using (2.6) and the linearity of T:

$$\sum_{i=1}^{n} \lambda_i^{m-1} \beta_i T e_i = \sum_{i=1}^{n} \lambda_i^m \beta_i e_i \tag{2.11}$$

Hence:

$$Te_i = \lambda_i e_i \tag{2.12}$$

i.e. the λ_i's are also eigenvalues of the operator T that acts in a finite (and small) dimensional space. Let:

$$e_i = \sum_{j=1}^{n} g_{ij} v_j \quad 1 \leq i,j \leq n \tag{2.13}$$

be an expansion of the e_i's in terms of the v_j's. Then:

$$Te_i = \sum_{j=1}^{n} g_{ij} T v_j \tag{2.14}$$

Define now the matrix **D** by:

$$Tv_i = D_{ij} v_j \quad 1 \leq i,j \leq n \tag{2.15}$$

where the (Einstein) summation convention over repeated indices is assumed. It follows from (2.10) that $D_{ij} = \delta_{j,i+1}$ for $1 \leq i \leq n-1$ and from (2.9) that $D_{n,j} = \gamma_j$. D is just a representation of the T transformation.

Upon substituting (2.12), (2.13) in the left hand side of (2.14), and (2.15) in the right hand side of (2.14) we obtain:

$$\lambda_i g_{ik} v_k = g_{ij} D_{jk} v_k \tag{2.16}$$

Hence:

$$\lambda_i g_{ik} = g_{ij} D_{jk} \tag{2.17}$$

i.e. the row $g_{i,k}$ (i fixed) is a left eigenvector of the matrix **D** corresponding to the eigenvalue λ_i. Using the definition of the matrix **D** we can now proceed to find λ_i and $g_{i,k}$. Inspection of the matrix **D** for small n leads to the following result, which can be easily checked by substitution into (2.17):

$$g_{i,k} = \sum_{m=1}^{k} \frac{\gamma_m}{\lambda_i^{k+1-m}} \tag{2.18}$$

and

$$\sum_{m=1}^{n} \frac{\gamma_m}{\lambda_i^{n+1-m}} = 1 \tag{2.19}$$

Eqn. (2.19) can also be rewritten as:

$$\sum_{m=1}^{n} \gamma_m \lambda_i^{m-1} - \lambda_i^n = 0 \tag{2.20}$$

Consequently, given the values of γ_m, one can use (2.20) to compute the spectrum and (2.18) and (2.13) to compute the eigenvectors. It only remains to compute the values of γ_m. To this end we perform an orthogonalization of the vectors $v_1, \ldots, v_n$

leading to the orthonormal set $w_1 \cdots w_n$ such that:

$$w_i = \sum_{j=1}^{i} c_{ij} v_j \tag{2.21}$$

where the definition of the matrix c is obvious. Hence:

$$v_{n+1} = \sum_{k=1}^{n} (v_{n+1} \cdot w_k^{\dagger}) w_k \tag{2.22}$$

and using (2.21)

$$v_{n+1} = \sum_{m=1}^{n} \sum_{k=1}^{n} (v_{n+1} \cdot w_k^{\dagger}) c_{km}^{-1} v_m \tag{2.23}$$

Hence:

$$\gamma_m = \sum_{k=1}^{n} (v_{n+1} \cdot w_k^{\dagger}) c_{km}^{-1} \tag{2.24}$$

Eqns. (2.18), (2.20) and (2.24) constitute the sought solution for the eigenvalues and the eigenvectors.

All of the above formalism can be easily modified when the vectors v_m are defined by:

$$v_m = e^{(m-1)M\tau} \hat{v}(t) \tag{2.25}$$

where τ is a chosen time interval. In this case, λ_i in eqns. (2.6) — (2.7) is to be replaced by $e^{\lambda_i \tau}$. This choice is especially convenient when the matrix **M** actually represents a differential operator (such as the Orr-Sommerfeld operator). In such a case the vectors v_m can be obtained by solving $\dot{x} = Mx$ for $0 \leq t \leq \tau$ with v_{m-1} as the initial condition, i.e. solving an initial value problem for the differential operator.

In this way one avoids storing a large matrix $\mathbf{M}$ representing the differential operator.

Finally, we note that in the above procedure the coefficients of e_i for increasing values of i are increasingly damped by exponential factors. Moreover, the above method should work, strictly speaking, only for cases of nondegenerate spectra. A remedy to both of these problems is provided by a direct construction of the orthogonal vectors $\mathbf{w}_i$ [cf. (2.3) and (2.21)]. This is done by defining:

$$\mathbf{w}_1 = \hat{\mathbf{x}}(t) \tag{2.26}$$

and:

$$T\mathbf{w}_i = \sum_{k=1}^{i} a_{ik}\mathbf{w}_i + a_{i,i+1}\mathbf{w}_{i+1}; \quad 1 \le i \le n \tag{2.27}$$

where: $T\mathbf{w}_i = \mathbf{M}\mathbf{w}_i$ or $T\mathbf{w}_i = e^{\mathbf{M}\tau}\mathbf{w}_i$ corresponding to $Te_i = \lambda_i e_i$ or $Te_i = e^{\lambda_i \tau}e_i$, respectively (which are the two alternatives presented above). The advantage of computing the vectors $\mathbf{w}$ by the use of (2.27) is that it enhances the weight of nonleading eigenvectors (see Section III for details).

The net result of the formalism presented in this section is the construction of n linearly independent vectors which are spanned (to a good approximation) by the first n eigenvectors of $\mathbf{M}$. Then either equation (2.20) or a representation of the operator T in the basis defined by $\{\mathbf{w}_i;\ 1 \le i \le m\}$ can be used for the computation of the eigenvalues and eigenvectors. The resulting method is economical and simple to implement.

III. ANALYSIS OF THE METHOD

The present section is devoted to an analysis of the formalism developed in section II. More specifically, we investigate the error involved in using a finite number of vectors e_i (or w_i). In what follows we shall use the "exponential" version for the operator T: $Te_i = e^{\lambda_i \tau} e_i$.

Firstly, we wish to investigate the extent to which the (right) eigenvectors, $\{e_i\}$, of M are spanned by the vectors $\{w_i\}$ or by the vectors $\{v_i\}$. Then we propose to estimate the error involved in the computation of the eigenvalues.

We wish to show that the eigenvectors e_i can be written as follows:

$$e_i = \sum_{j=1}^{i} a_{ij} w_j + \sum_{j=i+1}^{R} a_{ij} e^{\lambda_{ji} t} e_j \tag{3.1}$$

where R is the rank of the matrix M, $\lambda_{ji} \equiv \lambda_j - \lambda_i$, t is the time of filtering or of initial integration and the coefficients a_{ij} are of order unity. The proof proceeds by induction. By construction:

$$x(t) = \sum_{i=1}^{R} e^{\lambda_i t} \alpha_i e_i \tag{3.2}$$

where α_i are coefficients [see eqn. (2.1)]. Consequently, w_1, being a normalized vector in the direction of $x(t)$, can be written as:

$$w_1 = N_1 \sum_{i=1}^{R} e^{\lambda_{i1} t} \alpha_i e_i \tag{3.3}$$

where the right side of eqn. (3.2) was divided by $e^{\lambda_1 t}$ and normalized. N_1 is a normalization factor. Hence:

$$\mathbf{e}_1 = \frac{1}{N_1\alpha_1}\mathbf{w}_1 - \frac{1}{\alpha_1}\sum_{i=2}^{R} e^{\lambda_{i1}t}\alpha_i\mathbf{e}_i \tag{3.4}$$

which shows that $\mathbf{e}_1$ indeed has the expected form (with $a_{11} = \frac{1}{N_1\alpha_1}$ and $a_{1,i} = -\frac{\alpha_i}{\alpha_1}$ for $i \geq 2$). Next we prove that eqn. (3.2) is valid for $i = 2$. By definition:

$$\mathbf{w}_2 = N_2(\mathbf{v}_2 - (\mathbf{v}_2 \cdot \mathbf{w}_1^{\dagger})\mathbf{w}_1) \tag{3.5}$$

where N_2 is a normalization factor. Hence:

$$\frac{1}{N_2}\mathbf{w}_2 + (\mathbf{v}_2 \cdot \mathbf{w}_1^{\dagger})\mathbf{w}_1 = \alpha_1 e^{\lambda_1(t+\tau)}\mathbf{e}_1 + \alpha_2 e^{\lambda_2(t+\tau)}\mathbf{e}_2 + \sum_{i=3}^{R}\alpha_i e^{\lambda_i(t+\tau)}\mathbf{e}_i \tag{3.6}$$

Substituting $\mathbf{e}_1$ from eqn. (3.4) in the right side of eqn. (3.6):

$$\frac{1}{N_2}\mathbf{w}_2 + (\mathbf{v}_2 \cdot \mathbf{w}_1^{\dagger})\mathbf{w}_1 = \frac{1}{N_1}e^{\lambda_1(t+\tau)}\mathbf{w}_1 - e^{\lambda_1(t+\tau)}e^{\lambda_{21}t}\alpha_2\mathbf{e}_2$$

$$- e^{\lambda_1(t+\tau)}\sum_{i=3}^{R} e^{\lambda_{i1}t}\alpha_i\mathbf{e}_i + \alpha_2 e^{\lambda_2(t+\tau)}\mathbf{e}_2 + \sum_{i=3}^{R}\alpha_i e^{\lambda_i(t+\tau)}\mathbf{e}_i \tag{3.7}$$

By solving for $\mathbf{e}_2$ in eqn. (3.7) it is easy to see that formula (3.1) is correct for $i = 2$. Assume next that eqn. (3.1) has been proven for $1 \leq i \leq m$. Define:

$$A_{ij} = \begin{cases} a_{ij} & 1 \leq j \leq i \leq m \\ 0 & m \geq j > i \geq 1 \end{cases} \tag{3.8}$$

$$B_{ij} = \begin{cases} a_{ij}e^{\lambda_{ji}t} & m \geq j > i \geq 1 \\ 0 & 1 \leq j \leq i \leq m \end{cases} \tag{3.9}$$

Both $\mathbf{A}$ and $\mathbf{B}$ are square matrices of rank m. Define also:

$$C_{ij} = \begin{cases} a_{ij}e^{\lambda_{ji}t} & j \geq m+1;\ 1 \leq i \leq m \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

By the induction assumption [and using definitions (3-8, 9, 10) to rewrite eq. (3.1)] we have:

$$e_i = \sum_{j=1}^{m} A_{ij} w_j + \sum_{j=1}^{m} B_{ij} e_j + \sum_{j=1}^{R} C_{ij} e_j \tag{3.11}$$

for $1 \leq i \leq m$. Hence:

$$e_i = (\mathbf{I} - \mathbf{B})_{ik}^{-1} (A_{kj} w_j + C_{kj} e_j) \tag{3.12}$$

where the summation convention is assumed. For every integer r:

$$(\mathbf{B}^r)_{i,k} = a_{i,i_1} a_{i_1,i_2} \cdots a_{i_{r-1},k} e^{\lambda_{i_1,i}t + \lambda_{i_2,i_1}t + \cdots \lambda_{k,i_{r-1}}t} \tag{3.13}$$

or

$$B^r_{i,k} \propto e^{\lambda_{k,i}t} \tag{3.14}$$

Consequently:

$$(\mathbf{I} - \mathbf{B})_{ik}^{-1} = \delta_{ik} + h_{ik} e^{\lambda_{k,i}t} \tag{3.15}$$

where h_{ik} is O(1) and $h_{ik} = 0$ for $i \geq k$ or $k > m$ [see eqn. (3.9)]. Substituting eqn. (3.15) in (3.12) we obtain:

$$e_i = A_{ij} w_j + C_{ij} e_j + h_{ik} e^{\lambda_{k,i}t} A_{kj} w_j + h_{ik} e^{\lambda_{k,i}t} C_{kj} e_j; \quad i \leq i \leq m \tag{3.17}$$

The first and third terms in the right side of eqn. (3.17) are linear combinations of $\{w_k;\ 1 \leq k \leq m\}$. The second term is, using eqn. (3.10):

$$\sum_{j=m+1}^{R} a_{ij} e^{\lambda_j t} e_j \tag{3.18}$$

The fourth term is:

$$\sum_{k=i+1}^{R} h_{ik} e^{\lambda_k t} \sum_{j=m+1}^{R} a_{kj} e^{\lambda_j t} e_j = \sum_{j=m+1}^{R} \left(\sum_{k=i+1}^{R} h_{ik} a_{kj} \right) e^{\lambda_j t} e_j \tag{3.19}$$

Consequently, by separating the terms containing $\{w_j; 1 \leq j \leq i\}$ and those which are superpositions of $\{w_j; i+1 \leq j \leq m\}$ in eqn. (3.17) we obtain:

$$e_i = \sum_{j=1}^{i} \gamma_{ij} w_j + \sum_{k=i+1}^{m} \sum_{j=i+1}^{k} e^{\lambda_k t} h_{ik} A_{kj} w_j + \sum_{j=m+1}^{R} r_{ij} e^{\lambda_j t} e_j \tag{3.20}$$

where γ_{ij} and r_{ij} are $O(1)$ quantities. Note that we used the fact that $h_{ik} = 0$ for $i \geq k$ or $k > m$. It follows from eq. (3.20) that

$$e_i = \sum_{j=1}^{i} \gamma_{ij} w_j + \sum_{j=i+1}^{m} \sigma_{ij} e^{\lambda_j t} w_j + \sum_{j=m+1}^{R} r_{ij} e^{\lambda_j t} e_j \tag{3.21}$$

for $1 \leq i \leq m$, where the definition of σ_{ij} is obvious. Next we use eqn. (3.21) to complete the induction process. Eqn. (3.21) itself follows from the induction assumption for $1 < j \leq m$.

By its definition:

$$w_{m+1} = N_{m+1}\left(v_{m+1} - \sum_{i=1}^{m} (v_{m+1} \cdot w_i^{+}) w_i\right) \tag{3.22}$$

where N_{m+1} is a normalization factor. Consequently:

$$v_{m+1} = \sum_{i=1}^{m+1} q^{(1)}_{m+1,i} w_i \tag{3.23}$$

where $q^{(1)}_{m+1,i}$ are O(1) numbers. Using (2.4b) and (2.25):

$$\sum_{i=1}^{m+1} q^{(1)}_{m+1,i}\mathbf{w}_i = \sum_{i=1}^{m} \beta_i e^{\lambda_i(m\tau)}\mathbf{e}_i + \sum_{i=m+1}^{R} \beta_i e^{\lambda_i(m\tau)}\mathbf{e}_i \tag{3.24}$$

By assumption, the vectors $\mathbf{e}_i$; $1 \leq i \leq m$ can be expressed by eq.(3.21). Substituting eqn. (3.21) in eqn. (3.24) we obtain, after a rearrangement of terms:

$$\sum_{i=1}^{m+1} q^{(2)}_{m+1,i}\mathbf{w}_i = \sum_{i=1}^{m} \beta_i e^{\lambda_i(m\tau)} \sum_{j=m+1}^{R} r_{ij} e^{\lambda_{ji}t}\mathbf{e}_j + \sum_{i=m+1}^{R} \beta_i e^{\lambda_i(m\tau)}\mathbf{e}_i \tag{3.25}$$

where $q^{(2)}_{m+1,i}$ are the new coefficients of the $\mathbf{w}_i$ terms. Solving for $\mathbf{e}_{m+1}$ in eq.(3.25) we obtain (using eq. (2.4c) for β_i):

$$\mathbf{e}_{m+1} = \sum_{i=1}^{m} q_{m+1,i}\mathbf{w}_i + \sum_{k=m+2}^{R} s_{m+1,k} e^{\lambda_{k,m+1}t}\mathbf{e}_k \tag{3.26}$$

where $q_{m+1,i}$ and $s_{m+1,k}$ are O(1) quantities (assuming $e^{\lambda_{ji}t}$ is O(1)). This completes the induction. Consequently eqn. (3.1) and eqn. (3.21) are correct for all $i \geq 1$.

It follows from eq. (3.21) that the error involved in the assumption that $\mathbf{e}_i$ is a combination of $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m$ is $O(e^{\lambda_{m+1,i}t})$, which means the choice of the size of the subspace of $\{\mathbf{w}_i;\ 1 \leq i \leq m\}$ should be such as to have $|\mathrm{Re}(\lambda_{m+1} - \lambda_i)t| >> 1$ for $1 \leq i \leq r$ if r correct eigenvectors are wanted. Notice that no gap in the spectrum is necessary for this estimate or the proposed method to be valid. Some of the eigenvectors (i.e. those for which $|\mathrm{Re}(\lambda_{m+1} - \lambda_i)t|$ is not large enough) will not be well approximated by the procedure. In such a case an appropriate increase of the value of m will ensure a good approximation for $\mathbf{e}_i$. A similar statement will be shown below to be true for the corresponding eigenvalues. Before we do that, we present a second result approximating the error involved in expressing the eigenvectors e_i in terms of the v_m's [see eqns. (2.8), (2.9)].

We wish to show that for each $1 \leq i \leq R$ there is a linear combination of $\{v_i; 1 \leq i \leq R\}$, which we denote by u_i and which satisfies:

$$u_i = e_i + \sum_{j=i+1}^{R} K_{ij} e^{\lambda_{ji} t} e_j \tag{3.27}$$

where K_{ij} are O(1). As before this statement may be proven by straightforward induction. Here we shall only sketch a proof. For i = 1 define

$$u_1 = \frac{1}{\alpha_1 e^{\lambda_1 t}} \sum_{j=1}^{R} \alpha_j e^{\lambda_j t} e_j \tag{3.28}$$

Hence:

$$u_1 = e_1 + \sum_{j=2}^{R} \frac{\alpha_j}{\alpha_1} e^{\lambda_{j1} t} e_j \tag{3.29}$$

Next define $v_k^{(1)}$ as v_k divided by $\alpha_1 e^{\lambda_1 (t+(k-1)\tau)}$:

$$v_k^{(1)} = e_1 + \sum_{j=2}^{R} \frac{\alpha_j}{\alpha_1} e^{\lambda_{j1}(t+(k-1)\tau)} \tag{3.30}$$

Hence:

$$v_k^{(1)} - u_1 = \sum_{j=2}^{R} \frac{\alpha_j}{\alpha_1} e^{\lambda_{j1} t} (e^{\lambda_{j1}(k-1)\tau} - 1) e_j \tag{3.31}$$

Dividing $v_k^{(1)} - u_1$ by $\frac{\alpha_2}{\alpha_1} e^{\lambda_{21} t}$ we obtain:

$$v_k^{(2)} = \sum_{j=2}^{R} \frac{\alpha_j}{\alpha_2} e^{\lambda_{j2}t}(e^{\lambda_{j1}(k-1)\tau} - 1)e_j \tag{3.32}$$

Next define:

$$u_2 \equiv v_2^{(2)}/(e^{\lambda_{21}t} - 1)$$

or:

$$u_2 = e_2 + \sum_{j=3}^{R} \frac{\alpha_j}{\alpha_2} \frac{e^{\lambda_{j2}t}(e^{\lambda_{j1}\tau} - 1)}{e^{\lambda_{21}\tau} - 1} e_j \tag{3.33}$$

A similar procedure of Gauss elimination steps leads to:

$$u_3 = e_3 + \sum_{j=4}^{R} \frac{\alpha_j}{\alpha_3} e^{\lambda_{j3}t} \frac{\dfrac{e^{2\lambda_{j1}\tau} - 1}{e^{2\lambda_{21}\tau} - 1} - \dfrac{e^{\lambda_{j1}\tau} - 1}{e^{\lambda_{21}\tau} - 1}}{\dfrac{e^{2\lambda_{31}\tau} - 1}{e^{2\lambda_{21}\tau} - 1} - \dfrac{e^{\lambda_{31}\tau} - 1}{e^{\lambda_{21}\tau} - 1}} e_j \tag{3.34}$$

The continuation of this process obviously leads to the desired equation (3.27). It is now easy to see the e_i can be expressed in terms of $\{v_1, \cdots v_m\}$ with an error of $O(e^{\lambda_{m+1,i}t})$. Notice that when τ itself is large the coefficient of e_j in eqn. (3.34) tends to zero, which shows that one may obtain adequate approximations for large values of τ and (even) short values of t.

Finally we turn to estimating the error in the eigenvalues, when computed using a finite number (n) of vectors. To this end define a matrix U by:

$$U_{ij} = w_j^{+} T w_i \qquad 1 \leq i,j \leq n \tag{3.35}$$

U is the reduction (or projection) of the operator T to the finite dimensional subspace spanned by the v's (or w's). Let ϕ be an eigenvector of U, with a corresponding eigenvalue Λ:

$$\sum_{j=1}^{m} U_{ij}\phi_j = \Lambda\phi_i \tag{3.36}$$

Let:

$$e_i = \sum_{j=1}^{R} b_{ij}w_j \qquad 1 \leq i \leq R \tag{3.37}$$

Then, since $Te_i = e^{\lambda_i\tau}e_i$, we have:

$$\sum_{j=1}^{R} b_{ij}Tw_j = e^{\lambda_i\tau}\sum_{j=1}^{R} b_{ij}w_j \tag{3.38}$$

Hence:

$$Tw_i = \sum_{j,k=1}^{R} b_{ik}^{-1}e^{\lambda_k\tau}b_{kj}w_j \tag{3.39}$$

and, using eqn. (3.35):

$$U_{ij} = \sum_{k=1}^{R} b_{ik}^{-1}e^{\lambda_k\tau}b_{kj} \tag{3.40}$$

Hence, from eqns. (3.36) and (3.40):

$$\sum_{j=1}^{m}\sum_{k=1}^{R} b_{ik}^{-1}e^{\lambda_k\tau}b_{kj}\phi_j = \Lambda\phi_i \tag{3.41}$$

Define:

$$\psi_k = \sum_{j=1}^{m} b_{kj}\phi_j \tag{3.42}$$

Hence

$$\sum_{i=1}^{m}\sum_{k=1}^{R} b_{ri}b_{ik}^{-1}e^{\lambda_k\tau}\psi_k = \Lambda\psi_r \tag{3.43}$$

Since $\sum_{i=1}^{R} b_{ri}b_{ik}^{-1} = \delta_{rk}$, we have:

$$e^{\lambda_r t}\psi_r + \sum_{k=1}^{R}\sum_{i=m+1}^{R} b_{ri}\, b_{ik}^{-1}e^{\lambda_k\tau}\psi_k = \Lambda\psi_r \tag{3.44}$$

By comparing eqn. (3.37) and eqn. (3.1) or (3.21) we observe that:

$$b_{ij} = \begin{cases} a_{ij} & j \leq i \\ \overline{a}_{ij}e^{\lambda_{j,}t} & j > i \end{cases} \tag{3.45}$$

where the definition of $\overline{a}_{ij}$ is obvious. Using eqns. (3.8), (3.9) with $m = R$ and a_{ij} replaced by $\overline{a}_{ij}$ in eq. (3.9) we see that:

$$\mathbf{b} = \mathbf{A} + \mathbf{B} \tag{3.46}$$

Hence:

$$\mathbf{b}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\cdots \tag{3.47}$$

Since $A_{ki}^{-1} = 0$ for $k < i$, by definition, we have to estimate only terms containing elements of the matrix **B** in (3.47). In this case:

$$(A^{-1}BA^{-1})_{ik} = A_{ii_1}^{-1}B_{i_1i_2}A_{i_2k}^{-1} \tag{3.48}$$

or:

$$(A^{-1}BA^{-1})_{ik} = \sum_{i_1 \leq i} A^{-1}_{ii_1} \sum_{i_1 < i_2} \overline{a}_{i_1 i_2} e^{\lambda_{i_2 i_1} t} A^{-1}_{i_2 k} \tag{3.49}$$

The largest contribution of $e^{\lambda_{i_2 i_1} t}$ which is possible under the constraints $k \leq i_2 > i_1 \leq i$ is for $i_2 = k$ and $i_1 = i$. A similar analysis holds for all terms in the sum eq. (3.47). Hence:

$$b^{-1}_{ik} = e^{\lambda_{ki} t} d_{ik} \quad \text{for } i < k \tag{3.50}$$

where d_{ik} are O(1) numbers. When $i \geq k$, A^{-1}_{ik} is O(1) itself and we denote it by d_{ik}. Hence, from (3.45):

$$e^{\lambda_r \tau}\psi_r + \sum_{i=m+1}^{R} \sum_{k=1}^{i} b_{ri} d_{ik} e^{\lambda_k \tau}\psi_k + \sum_{i=m+1}^{R} \sum_{k=i+1}^{R} b_{ri} d_{ik} e^{\lambda_k \tau} e^{\lambda_{ki} t}\psi_k = \Lambda\psi_r \tag{3.51}$$

Using (3.45) again (to substitute for b_{ri}):

$$e^{\lambda_r \tau}\psi_r + \sum_{i=m+1}^{R} \sum_{k=1}^{i} \overline{a}_{ri} e^{\lambda_{ir} t} d_{ik} e^{\lambda_k \tau}\psi_k +$$

$$\sum_{i=m+1}^{R} \sum_{k=i+1}^{R} \overline{a}_{ri} e^{\lambda_{ir} t} d_{ik} e^{\lambda_k \tau} e^{\lambda_{ki} t}\psi_k = \Lambda\psi_r \tag{3.52}$$

For $|Re(\lambda_{m+1} - \lambda_r)t| >> 1$ the leading order terms in the double sums of eqn. (3.52) are $e^{\lambda_{m+1,r} t}\overline{a}_{r,m+1} \sum_{k=1}^{m+1} d_{m+1,k} e^{\lambda_k \tau}\psi_k$ and $e^{\lambda_{m+2,r} t}\overline{a}_{r,m+1} d_{m+1,m+2} e^{\lambda_{m+2} \tau}\psi_{m+2}$, respectively. Hence:

$$(e^{\lambda_r \tau} - \Lambda)\psi_r + e^{\lambda_{m+1,r} t}\overline{a}_{r,m+1}\left(\sum_{k=1}^{m+1} d_{m+1,k} e^{\lambda_k \tau}\psi_k + e^{\lambda_{m+2,m+1} t} d_{m+1,m+2} e^{\lambda_{m+2} \tau}\psi_{m+2}\right) \simeq 0 \tag{3.53}$$

To lowest order (i.e. neglecting $e^{\lambda_{m+1,r} \tau}$) an eigensolution satisfies: $\Lambda = e^{\lambda_r \tau}$ and $\psi_j = \delta_{jr}$. The next perturbative correction yields:

$$\Lambda = e^{\lambda_r \tau} + e^{\lambda_{m+1,r}t}\bar{a}_{r,m+1}d_{m+1,r}e^{\lambda_r \tau} \tag{3.54}$$

which indeed shows that for a large enough filtration time, the error in the eigenvalues is exponentially small provided $\mathrm{Re}\lambda_{m+1}\cdot t$ is well separated from (or much smaller than) $\mathrm{Re}\lambda_r\cdot t$. Thus no gap in the spectrum is necessary to obtain excellent values for the leading eigenvalues. Obviously an increase in m will make the term $e^{\lambda_{m+1,r}t}$ as small as needed.

IV. THE ALGORITHMS

In this section we present detailed algorithms for the computation of $\lambda_1, \cdots, \lambda_n$ and $e_1, \cdots, e_n$. The algorithm is described in steps:

(1) Choose an initial vector x_0. It does not matter if x_0 is independent of e_1 or other relevant eigenvectors. They will be introduced in step (2) by round-off errors during the filtration process.

(2) Solve the equation $\dot{x} = Mx$, with x_0 as initial condition, up to a time t. Normalize the resulting $x(t)$: $x_1 = \frac{x(t)}{|x(t)|}$. Use now x_1 as an initial condition and solve for a time θ, obtaining $x_1(\theta)$. Repeat this process r times to obtain:

$$x_r = \frac{x(r\theta)}{|x(r\theta)|} \tag{4.1}$$

The reason one normalizes after each time t is to avoid dealing with large numbers (since $|x(t)| \propto e^{\lambda_1 t}$ for large times). The choice of r is explained below. Since it is not clear apriori whether x_0 is independent of e_1, e_2 etc., it is advisable to use low accuracy computation in the first few iterations. In this way the weight of e_1, e_2, etc., will be amplified.

(3) Compute $v_k = M^{k-1}x_n$, $1 \leq k \leq m+1$. Orthonormalize $v_1, v_2, \cdots v_m$ [using Householder transformations (Wilkinson, 1965), for example]. The resulting orthonormal vectors are denoted by $w_1, w_2, \cdots w_m$. Following each step of the orthonormalization pro-

cedure use the test of step (4) to make sure $v_1, \cdots v_m$ are independent. The matrix $\mathbf{c}$, defined by $w_i = \sum_{j=1}^{i} c_{ij} v_j$, that results from the orthonormalization procedure should be kept.

(4) Test whether $\mathbf{v}_{m+1}$ is spanned by $\mathbf{v}_1, \cdots \mathbf{v}_m$. To do this compute:

$$\|\mathbf{E}\| = \left|\mathbf{v}_{n+1} - \sum_{i=1}^{m} (\mathbf{v}_{m+1} \cdot \mathbf{w}_i)\mathbf{w}_i\right| / |\mathbf{v}_{m+1}| \tag{4.2}$$

If the error E is larger than desired, increase m. If this process results in too large a value for m, go back to step (2) and do several additional iterations. Then repeat steps (3) and (4) until E is smaller than the desired accuracy.

(5) Once the matrix $\mathbf{c}$ and the vectors $\mathbf{v}_k$ and $\mathbf{w}_k$ are known, use (2.24) to find the γ_k's. Alternatively solve the least square problem minimizing the expression $\|\mathbf{v}_{m+1} - \sum_{i=1}^{m} \gamma_i \mathbf{v}_i\|^2$. Standard least square routines may be used to solve for the γ_i's directly.

(6) Use (2.20) to find the spectrum.

(7) Use (2.18) to find the eigenvectors.

It should be stressed that in step (4) the test should be performed for m = 2, 3, etc., up to a desired m so as to make sure that $\mathbf{v}_1, \cdots \mathbf{v}_m$ are indeed independent. It may happen that the test in step (4) results in a value of m which is smaller than its desired value. In this case, either use a shorter integration time t or the procedure described below. If spurious eigenvalues appear, they will be mdependent. A comparison of the eigenvalues for different values of m enables the identification of the nonspurious eigenvalues.

The reason the procedure described above may converge for a value of m (say m_1) which is smaller than the desired (number of eigenvalues) m can be a gap in the spectrum at m_1, namely:

$$\mathrm{Re}\lambda_1 > \mathrm{Re}\lambda_2 \cdots \mathrm{Re}\lambda_{m_1} >> \mathrm{Re}\lambda_{m_1+1} > \cdots \tag{4.3}$$

In this case, $\lambda_1, \lambda_2, \cdots \lambda_{m_1}$ and the corresponding eigenvectors are obtained to an exponential accuracy which is demonstrated in the next section. Therefore it is possible to obtain the m_1 left eigenvectors $\mathbf{e}_i^L$ corresponding to $\mathbf{e}_i$, $1 \leq i \leq m_1$. These vectors satisfy:

$$\mathbf{e}_i^L \cdot \mathbf{e}_j = \delta_{ij} \tag{4.4}$$

Expanding $\mathbf{e}_j$ in the orthonormal set $\mathbf{w}$ (see eqn. (2.21)):

$$(\mathbf{e}_i^L \cdot \mathbf{w}_k)(\mathbf{w}_k^+ \mathbf{e}_j) = \delta_{ij} \tag{4.5}$$

Thus $(\mathbf{e}_i^L \cdot \mathbf{w}_k)$ is the inverse of the (known) matrix A (of rank m_1):

$$Q_{kj} = (\mathbf{w}_k^{\dagger} \mathbf{e}_j) \tag{4.6}$$

Consequently:

$$\mathbf{e}_i^L = \sum_{j=1}^{m_1} Q_{ij}^{-1} \cdot \mathbf{w}_j^{\dagger} \tag{4.7}$$

Consider now a vector $\mathbf{x}_0$ that is chosen to be orthogonal to $\mathbf{e}_i^L$, $1 \leq i \leq m$. Now perform step (2) with the modification that after each time θ, $x_r(t)$ is orthogonalized to $\mathbf{w}_1, \cdots \mathbf{w}_{m_1}$. The resulting vector, after a time t, is:

$$y(t) = \frac{1}{\rho}\left[e^{Mt}x_0 - \sum_{i=1}^{m_1}[(e^{Mt}x_0)\cdot w_i^{+}]w_i\right] \tag{4.8}$$

where ρ is a normalization factor introduced in order to ensure $|y| = 1$. Since x has been orthonormalized to w_i it follows that:

$$x = \sum_{i=1}^{m_1}\varepsilon_i e_i + \sum_{i>m_1}\alpha_i e_i \tag{4.9}$$

where ε_i are exponentially small quantities. Substituting (4.9) into (4.8) results in:

$$y(t) = \frac{1}{\rho}\sum_{i=1}^{m_1}e^{\lambda_i t}\varepsilon_i[e_i - (e_i w_j^{+})w_j] + \frac{1}{\rho}\sum_{i>m_1}e^{\lambda_i t}\alpha_i[e_i - (e_i\cdot w_j^{+})w_j] \tag{4.10}$$

The difference $[e_i - (e_i w_j^{+})w_j]$ is exponentially small for the exact e_i's. So is $(e_i w_j^{+})$ for $i > m_1$. Thus, for t not too large: $y(t) \propto \sum_{i>m_1} e^{\lambda_i t}\alpha_i$ and the regular procedure can be applied to y(t) to find λ_{m_1}, λ_{m_1+1}, etc. If λ_{m_1} is close to λ_{m_1+1} a relatively large time t in (4.10) may be necessary. In this case a better accuracy for e_i, $1 \le i \le m_1$ should be obtained first. Methods to do so will be given in a future publication.

Another version of the suggested algorithm which is highly suitable when no gap in the spectrum exists is based on eqns. (2.25) — (2.26). Step (2) involves then an orthogonalization after each integration for a time of τ. The projection of the operator T (in the subspace spanned by the w's) can then be expressed as shown in Section III [eqn. (3.35)]. The small resulting $m\times m$ matrix can be diagonalized by LR/QR methods (Wilkinson, 1965) (we have used EIGCG1, an EISPACK routine) yielding exponentially accurate results.

We summarize this section by mentioning that the above method can be efficiently used both when the available computer has a small memory (then increase T_f) and when it has a large one (then more vectors and eigenvalues can be computed).

V. IMPLEMENTATION OF THE METHOD: THE ORR-SOMMERFELD EQUATION.

In this section we present an application of the methods described above to a problem in hydrodynamical stability[2]: the spectrum of the Orr-Sommerfeld equation for channel flow. This equation has an infinite number of degrees of freedom and there is no "gap" in its spectrum. Thus it is a good test case for our methods. Moreover, the existence[11] of previously computed, highly accurate values of the spectrum enables us to perform a comparison of our results with the known spectrum.

The Orr-Sommerfeld equation reads:

$$(U-c)(D^2-\alpha^2)\psi - U''\psi = \frac{1}{i\alpha R}(D^2-\alpha^2)^2\psi \tag{5.1}$$

$$\psi(x,y,t) = \psi(y)\exp[i\alpha(x-ct)]$$

where $U(y)$ for $-1 \leq y \leq 1$, is the basic velocity profile, $\psi(x,y,t)$ is the perturbation streamfunction, R the Reynolds number and D the cross-stream derivative $\frac{d}{dy}$. The real number α represents the wavenumber of the streamwise periodic perturbation and c is the sought complex eigenvalue. The real part of the eigenvalue c_r is the phase speed of the perturbation. The growth rate of the perturbation is $\exp(\alpha c_i t)$ where $c_i = \text{Im } c$. Thus, a perturbation ψ is stable if $c_i < 0$ and unstable if $c_i > 0$. The boundary conditions on eq.(5-1) are:

$$\psi(\pm 1) = D\psi(\pm 1) = 0 \tag{5.2}$$

The above is an eigenvalue problem. It is transforme[illegible] a time depen-dent problem (from which it is actually derived) by redefining the [illegible] stream function ψ as:

$$\psi(x,y,t) = \psi(y,t)e^{i\alpha x}$$

even though the equation is separable in time. We thus solve the following initial value problem:

$$\partial_t(D^2-\alpha^2)\psi + i\alpha U(D^2-\alpha^2)\psi - i\alpha U''\psi = R^{-1}(D^2-\alpha^2)^2\psi$$

$$\psi(\pm 1) = D\psi(\pm 1) = 0 \tag{5.2a,b}$$

It is convenient to define:

$$\zeta = (D^2-\alpha^2)\psi \tag{5.3}$$

$$\zeta = \rho + i\lambda \tag{5.4a}$$

$$\psi = \phi + i\chi \tag{5.4b}$$

to obtain the coupled real system:

$$\partial_t\rho - \alpha U\lambda + \alpha U''\chi = R^{-1}(D^2-\alpha^2)\rho \tag{5.5a}$$

$$\partial_t\lambda + \alpha U\rho - \alpha U''\phi = R^{-1}(D^2-\alpha^2)\lambda \tag{5.5b}$$

$$(D^2-\alpha^2)\phi = \rho \tag{5.5c}$$

$$(D^2-\alpha^2)\chi = \lambda \tag{5.5d}$$

with boundary conditions:

$$\phi(\pm 1) = \chi(\pm 1) = D\phi(\pm 1) = D\chi(\pm 1) = 0 \tag{5.6}$$

The real system may be formally written as $\dot{\psi} = M\psi$, M being an integro-differential operator.

These equations have been solved by the Chebycheff pseudo-spectral technique (see the appendix for details). An initial filtering time T_f was used to produce a function $\psi(y,t)$, which is spanned by a relatively small number of eigenvectors corresponding to leading eigenvalues $Re(-i\alpha c)$ or the fastest growing modes, as has been explained previously.

In all computations results have been presented for the parameter values $\alpha = 1.0$ and $R = 10{,}000$. The base velocity profile U is $(1 - y^2)$.

The results of our computations using the methods developed are presented in five tables. A sixth table of eigenvalues from ref.11 is taken as standard and used for comparing the accuracy of our results In all cases we obtain several leading eigenvalues to a very high accuracy.

In our computations we vary the filtration time T_f, the sampling interval τ, the accuracy of the time integration as expressed by the discrete time-step size during filtration δt_f and during the sampling δt_τ in order to elucidate their effect on the accuracy of the computed spectra and the possible generation of spurious modes. We have also investigated the influence of the initial streamfunction $\psi(y,0)$ and of the number of eigenvalues determined on the accuracy of the results.

Three methods have been used for obtaining the eigenvalues, as explained below. Table I presents results of 2 runs with a sampling time/delay time of 15 and an initial filtration of 50. The number of correct eigenvalues obtained using eq.(2-20) is 5 when the number of vectors is 8 or 16. The reason for this is the fact that the filtration time is too large to produce more eigenvalues accurately. For the case n = 16 we also obtain two dominant spurious modes. This is due to the sensitivity of the roots of the characteristic polynomial to small inaccuracies in the coefficients. The dominant spurious eigenvalues may be easily identified by their appearance as n, the number of vectors is increased. The initial condition for table I is $(1-y^2)^2 + \sum_{i=1}^{8} \psi_i$, where ψ_i are eigenvectors corresponding to the fastest growing 8 eigenmodes.

Table II was obtained with an initial condition of $\psi = U^2(1+y)$. The factor $(1+y)$ is introduced to ensure that the initial condition has both an even and an odd part as U^2 is even in y. As the filtration time is increased from $T_f = 50$ to $T_f = 75$ the number of accurate eigenvalues rises to 5. The longer filtration time allows for a better damping of the decaying modes not of interest. It is possible that the cruder time step in the second case, $\delta t_f = 0.075$ as opposed to $\delta t_f = 0.05$ may introduce modes independent of the initial condition by way of numerical noise.

In Table III the vectors v_m were generated from the initial condition $U^2(1+0.71y)$ after a filtration of $T_f = 75$ and a delay interval of $\tau = 20$. However, instead of using eq.(2-20) for finding the eigenvalues, we have employed the LR algorithm as implemented in the EISPACK routine EIGCG1 to calculate the eigenvalues of the matrix directly. Four eigenvalues are obtained accurately both for n= 8 and n= 16 vectors.

Tables IV and V were produced using the method of orthogonalization described at the end of section II (see eq.(2-25,26)). The eigenvalues of the reduced matrix obtained were again computed using EIGCG1. In table IV an increase of T_f from 50 to 150 leads to an increase in the number of accurate eigenvalues from 3 to 5 for n = 8 vectors. The improved accuracy using this last method is obtained, as explained in section III, by the amplification of the weight of the non-leading eigenvectors in the orthogonalization procedure. Indeed, as seen in table V, the number of accurate eigenvalues when n= 32 vectors is used is 13. The method of orthogonalization produces 12 accurate eigenvalues for n=32 even when no filtration is invoked. In all cases we have obtained excellent accuracy for the leading eigenvalue. The subdominant eigenvalues corresponding to anti-symmetric modes tend to be somewhat more accurate than eigenvalues corresponding to symmetric modes.

VI. SUMMARY AND CONCLUSIONS

We have shown how one can obtain leading eigenvalues and eigenvectors for large asymmetric matrices using a relatively simple and economical numerical scheme. We have also shown how such a method can be applied when the matrix to be diagonalized is a differential operator. In the latter case our method does not require storing of an effective matrix, which represents the differential operator on a complete basis of expansion functions.

Three variants of the method were tested: a) A direct application of the formalism presented in section II. b) A construction of a set of vectors as explained in section II and the associated reduced matrix, followed by a diagonalization using standard eigenvalue routines for general matrices. c) Direct construction of a set of orthogonal vectors as at the end of section II followed by a diagonalization of the reduced matrix.

The third variant seems to be the superior method. The method was found to be very robust and did not require fine tuning to improve the accuracy of the calculated eigenvalues. Nor does it produce spurious dominant eigenvalues. The accuracy of the leading eigenvalues may be further increased by considering larger reduced matrices, that is by increasing the number of orthogonal eigenvectors. The errors involved in the proposed algorithm are analyzed in detail in section III and are in good agreement with our numerical results.

When a large matrix, of rank R, is considered the number of operations necessary to find its eigenvalues, using standard methods, scales like R^3. In our method, the number of operations is proportional to the rank R, the number of vectors used, m, and the time of filtration t_f, i.e. it is Rnt_f. The time of filtration depends on the spectrum, as explained in section III. If one wishes to obtain k correct eigenvalues one needs:

$$\left| e^{(\lambda_{m+1} - \lambda_k)t} \right| << 1$$

Thus the value of t_f is of the order of $\frac{1}{|\lambda_{m+1} - \lambda_k|}$.

Future applications of this method could include complex and non-Newtonian hydrodynamical stability problems, lattice eigenvalue problems and other systems leading to large asymmetric matrices for which the dominant eigenvalues are of interest and standard methods are either slowly converging or the memory requirements are prohibitively large. We believe that this method coupled to acceleration techniques may enable one to tackle many interesting eigenvalue problems.

APPENDIX

In this Appendix we present some details concerning the numerical solution of eqn. (5.4). We have employed a pseudospectral technique based on the expansion of ψ in Chebyshev polynomials, which gives a good resolution of boundary and critical layers. A third order Adams-Bashforth time-stepping scheme was used for an explicit evaluation of the advection (variable coefficient) terms and a second order Crank-Nicholson scheme for the linear diffusion terms. The time discretized equations read:

$$\frac{\rho^{(n+1)}-\rho^{(n)}}{\Delta t} - \alpha U[\frac{23}{12}\lambda^{(n)} - \frac{16}{12}\lambda^{(n-1)} + \frac{5}{12}\lambda^{(n-2)}]$$

$$+ \alpha U''[\frac{23}{12}\chi^{(n)} - \frac{16}{12}\chi^{(n-1)} + \frac{5}{12}\chi^{(n-2)}] = \frac{1}{2R}(D^2-\alpha^2)(\rho^{(n+1)}-\rho^{(n)}) \quad (A.1)$$

and

$$\frac{\lambda^{(n+1)}-\lambda^{(n)}}{\Delta t} + \alpha U[\frac{23}{12}\rho^{(n)} - \frac{16}{12}\rho^{(n-1)} + \frac{5}{12}\rho^{(n-2)}]$$

$$- \alpha U''[\frac{23}{12}\phi^{(n)} - \frac{16}{12}\phi^{(n-1)} + \frac{5}{12}\phi^{(n-2)}] = \frac{1}{2R}(D^2-\alpha^2)[\lambda^{(n+1)}+\lambda^{(n)}] \quad (A.2)$$

Eqns. (A.1) and (A.2) can be rearranged as follows:

$$(D^2-\alpha^2-\frac{2R}{\Delta t})\rho^{(n+1)} = F_1^{(n,n-1,n-2)}$$

$$(A.3)$$

$$(D^2-\alpha^2-\frac{2R}{\Delta t})\lambda^{(n+1)} = F_2^{(n,n-2,n-2)}$$

F_1 and F_2 being found from (A.1) and (A.2). The functions ϕ and χ satisfy [see eqns. (5.3), (5.5)]:

$$(D^2 - \alpha^2)\phi^{(n+1)} = \rho^{(n+1)}$$

$$(D^2 - \alpha^2)\chi^{(n+1)} = \lambda^{(n+1)} \tag{A.4}$$

Since the boundary conditions are on $\psi = \phi + i\chi$ alone, we have used the following Green function technique to satisfy these conditions. First, we find solutions to (A.3) satisfying $\rho^{(n+1)}(\pm 1) = \lambda^{(n+1)}(\pm 1) = 0$. We call these solutions $\rho_{hom}^{(n+1)}$ and $\lambda_{hom}^{(n+1)}$. These solutions are then substituted in the right hand side of eqns. (A.4) and solved using the (given) boundary conditions $\phi(\pm 1) = \chi(\pm 1) = 0$. We call these solution ϕ_{hom}^{n+1} and χ_{hom}^{n+1} respectively. They do not necessarily satisfy the Neumann boundary conditions $D\phi(\pm 1) = D\chi(\pm 1) = 0$. Next we solve the homogeneous equation:

$$(D^2 - \alpha^2 - \frac{2R}{\Delta t})\rho = 0; \;\; (D^2 - \alpha^2 - \frac{2R}{\Delta t})\lambda = 0 \tag{A.5}$$

using boundary conditions $\rho(1) = 1$, $\rho(-1) = 0$ and $\lambda(1) = 1$, $\lambda(-1) = 0$ respectively. These solutions are called ρ_+ and λ_+. Similarly we find ρ_- and λ_- which satisfy (A.5) with $\rho_-(1) = 0$, $\rho_-(0) = 1$ and $\lambda_-(1) = 0$, $\lambda_-(0) = 1$. Subsequently eqns. (A.4) are solved with ρ_+, ρ_- and λ_+, λ_- on the right side yielding $\phi_\pm$ and $\chi_\pm$ respectively (with boundary conditions $\phi_\pm(\pm 1) = \chi_\pm(\pm 1) = 0$ as required). The general solution for $\phi^{(n+1)}$ and $\chi^{(n+1)}$ can now be written:

$$\phi^{(n+1)} = \phi_{hom}^{(n+1)} + a_+\phi_+ + a_-\phi_-$$

$$\chi^{(n+1)} = \chi_{hom}^{(n+1)} + b_+\chi_+ + b_-\chi_- \tag{A.6}$$

The constants $a_\pm$ and $b_\pm$ are determined by imposing the Neumann boundary conditions $D\phi^{(n+1)}(\pm 1) = D\chi^{(n+1)}(\pm 1) = 0$ thus yielding a solution $\phi^{n+1} + i\chi^{n+1}$ which satisfies all four boundary conditions. The solutions $\phi_{\pm}$ and $\chi_{\pm}$ need only be computed once in a preprocessing step thus necessitating only two Poisson solvers per time step. The above leads to a very efficient time integration for the initial value problem.

REFERENCES

Arnoldi, W.E., Quart. Appl. Math., **9** 17, 1951.

Bennetin, G., Galgani, L., Giorgilli, A. and Strelcyn, J.M., Meccanica, **15** 9, 21, 1980.

Drazin, P.G. and Reid, W.H., The Theory of Hydrodynamic Stability, Cambridge University Press (New York, 1981).

Goldhirsch, I., Sulem, P.L. and Orszag, S.A., Stability and Lyapunov Stability of Dynamical Systems: A Differential Approach and A Numerical Method, Physica D, 1987 (in press).

Jenning, A. and Stewart, J., Jour. Inst. Math. Appl., **15** 351, 1975.

Lewis, J.G., Algorithms for Sparse Matrix Eigenvalue Problems, Ph.D. Thesis, Stanford Univ. Rep. 77-595, 1977.

Orszag, S.A., J. Fluid Mech., **50** 689-703, 1971.

Parlett, B.N., The Symmetric Eigenvalue Problem, Prentice-Hall, Englewood Cliffs, N.J., 1980.

Saad, Y., Linear Algebra and its Applications, **34** 269, 1980 (and references therein).

Shimada, I. and Nagashima, T., Prog. Theor. Phys., **61** 1605, 1979.

Stanley, H.E., Introduction to Phase-Transition and Critical Phenomena, Oxford University Press (New York, 1971).

Wilkinson, J.H., The Algebraic Eigenvalue Problem, Clarendon, Oxford University Press, 1965.

Table I

$T_f = 50 \; \tau = 15 \; \delta t_f = 0.05 \; \delta t_\tau = 0.005$

Identified Mode Number	Eigenvalue c_i
1	3.73967060-03
2	-3.51600679-02
3	-3.52076045-02
4	-5.08987828-02
5	-6.31504249-02
*	-8.72158010-02
*	-1.24740808-01
*	-1.32275383-01

$T_f = 50 \; \tau = 15 \; \delta t_f = 0.05 \; \delta t_\tau = 0.005$

Identified Mode Number	Eigenvalue c_i
1	3.73967060-03
*	4.58986340-04
*	3.40898714-04
*	-3.02658282-03
*	-1.49838951-02
*	-2.36492604-02
*	-2.93611665-02
2	-3.51423731-02
3	-3.52793842-02
4	-5.08971644-02
*	-1.09428509-01
*	-1.23518682-01
*	-1.48108621-01
*	-1.84825899-01

Table II

$T_f = 50$ $\tau = 20$ $\delta t_f = 0.05$ $\delta t_\tau = 0.005$ $T_f = 75$ $\tau = 20$ $\delta t_f = 0.075$ $\delta t_\tau = 0.005$

Identified Mode Number	Eigenvalue c_i	Identified Mode Number	Eigenvalue c_i
1	3.73967054-03	1	3.73967055-03
2	-3.51921344-02	2	-3.51630043-02
4	-5.09056673-02	3	-3.67725906-02
*	-5.52082215-02	4	-5.08973087-02
*	-7.13017236-02	5	-6.18479421-02
*	-1.32670178-01	*	-8.98244690-02
*	-1.47519731-01	*	-1.25561448-01
*	-1.64393611-01	*	-1.90016840-01

Table III

$T_f = 75$ $\tau = 20$ $\delta t_f = 0.075$ $\delta t_\tau = 0.005$		$T_f = 50$ $\tau = 20$ $\delta t_f = 0.05$ $\delta t_\tau = 0.005$	
Identified Mode Number	Eigenvalue c_i	Identified Mode Number	Eigenvalue c_i
1	3.73967061E-03	*	5.04217689E-03
2	-3.51313310E-02	1	3.73967060E-03
3	-3.77382799E-02	*	-9.48657987E-04
4	-5.08939102E-02	*	-1.15472857E-02
5	-6.08654660E-02	*	-1.55870014E-02
*	-8.94164908E-02	*	-2.30304183E-02
*	-1.23537573E-01	2	-3.40328560E-02
*	-1.83295676E-01	3	-3.51905276E-02
		4	-5.09188577E-02
		*	-5.12119074E-02
		5	-6.27014559E-02
		7	-9.93438231E-02
		9	-1.11875376E-01
		10	-1.11941416E-01
		*	-1.17783465E-01
		*	-1.70313262E-01

Table IV

$T_f = 50$ $\tau = 20$ $\delta t_f = 0.05$ $\delta t_\tau = 0.005$

$T_f = 75$ $\tau = 20$ $\delta t_f = 0.075$ $\delta t_\tau = 0.005$

$T_f = 150$ $\tau = 20$ $\delta t_f = 0.075$ $\delta t_\tau = 0.005$

Identified Mode Number	Eigenvalue c_i	Identified Mode Number	Eigenvalue c_i	Identified Mode Number	Eigenvalue c_i
1	3.73967060E-03	1	3.73967060E-03	1	3.73967060E-03
2	-3.51932116E-02	2	-3.51137133E-02	2	-3.51723718E-02
4	-5.09053371E-02	3	-3.74768214E-02	3	-3.51870632E-02
*	-5.48854885E-02	4	-5.08944600E-02	4	-5.08987622E-02
*	-7.36915731E-02	5	-6.07586848E-02	5	-6.30009355E-02
*	-1.38156654E-01	7	-9.17031242E-02	*	-1.37863162E-01
*	-1.41371117E-01	*	-1.23567838E-01	*	-1.43432356E-01
*	-1.58274564E-01	*	-1.82561894E-01	*	-1.74728812E-01

Table V

$T_f = 75$ $\tau = 20$ $\delta t_f = 0.075$ $\delta t_\tau = 0.005$		$T_f = 50$ $\tau = 20$ $\delta t_f = 0.05$ $\delta t_\tau = 0.005$		$T_f = 0$ $\tau = 20$ $\delta t_f =$ $\delta t_\tau = 0.005$	
Identified Mode Number	Eigenvalue c_i	Identified Mode Number	Eigenvalue c_i	Identified Mode Number	Eigenvalue c_i
1	3.73982608E-03	1	3.73967060E-03	1	3.73967061E-03
2	-3.51127796E-02	2	-3.51672225E-02	2	-3.51672225E-02
3	-3.51320836E-02	3	-3.51865287E-02	3	-3.51865287E-02
4	-5.08984842E-02	4	-5.08987305E-02	4	-5.08987305E-02
5	-6.31517698E-02	5	-6.32014442E-02	5	-6.32014442E-02
6	-6.32018420E-02	6	-6.32515219E-02	6	-6.32515219E-02
7	-9.11772541E-02	7	-9.12220344E-02	7	-9.12220344E-02
8	-9.12687323E-02	8	-9.13134832E-02	8	-9.13134832E-02
9	-1.19161983E-01	9	-1.19223288E-01	9	-1.19223290E-01
10	-1.19342042E-01	10	-1.19380224E-01	10	-1.19380226E-01
*	-1.21170936E-01	11	-1.24500810E-01	11	-1.24500810E-01
11	-1.24500425E-01	12	-1.38224635E-01	12	-1.38224635E-01
12	-1.38223897E-01	13	-1.45450671E-01	13	-1.45453318E-01
*	-1.44467829E-01	*	-1.51063411E-01	*	-1.51161943E-01
13	-1.45609239E-01	*	-1.58158173E-01	*	-1.58054756E-01
14	-1.46448075E-01	*	-1.60717285E-01	*	-1.60711699E-01
15	-1.75078430E-01	*	-1.61124155E-01	*	-1.61259736E-01
17	-1.81719355E-01	*	-1.64875002E-01	*	-1.64733124E-01
*	-1.97966129E-01	*	-1.67412781E-01	*	-1.67404319E-01
18	-2.04658289E-01	*	-1.72649311E-01	*	-1.72218483E-01
21	-2.07844699E-01	15	-1.75188825E-01	*	-1.74695007E-01
*	-2.20742079E-01	*	-1.78192328E-01	*	-1.80676443E-01
*	-2.22041721E-01	*	-1.81309666E-01	17	-1.82319221E-01
*	-2.67948177E-01	*	-1.82017191E-01	*	-1.83159305E-01
		17	-1.82776874E-01	*	-1.89780943E-01
		*	-1.92666139E-01	*	-1.95481111E-01
		*	-1.98326255E-01	*	-2.01652740E-01
		*	-2.02233301E-01	*	-2.02249437E-01
		*	-2.20500814E-01	*	-2.06222381E-01
		*	-2.36270762E-01	*	-2.34084842E-01
		*	-3.62463018E-01	*	-2.65098028E-01
		*	-3.88173344E-01	*	-3.11342750E-01

Table VI

Least stable eigenvalues for $\alpha = 1$, $R = 10000$

Mode Number	Eigenvalue c_i
1	+ 0.00373967
2	− 0.03516728
3	− 0.03518658
4	− 0.05089873
5	− 0.06320150
6	− 0.06325157
7	− 0.09122274
8	− 0.09131286
9	− 0.11923285
10	− 0.11937073
11	− 0.12450198
12	− 0.13822652
13	− 0.1472339
14	− 0.1474256
15	− 0.1752287
16	− 0.1754781
17	− 0.1828219
18	− 0.203221
19	− 0.203529
20	− 0.206465
21	− 0.208731
22	− 0.23119
23	− 0.23159
24	− 0.23882
25	− 0.25872
26	− 0.25969
27	− 0.25988
28	− 0.26511
29	− 0.26716
30	− 0.28551
31	− 0.28663
32	− 0.28765

END
11-87
DTIC